

KG-publicatie nr. 20

Statistische toetsen in Excel

Huub Everaert en Arie van Peet

KG-20

H. Everaert en A. van Peet
Statistische toetsen in Excel.

Correspondentie over deze KG-publicatie kunt u sturen naar:
E-mail: huub.everaert@hu.nl

KG-publicaties bevatten interne notities, verslagen en prepublicaties bestemd voor spreiding in kleine kring.
De verantwoordelijkheid voor de inhoud van een KG-publicatie berust bij de auteur(s). Uit KG-publicaties mag
alleen geciteerd worden met toestemming van de auteurs.

-
- nr.1 Everaert, H.A. en J.C. van der Wolf (2005).
 Behaviorally Challenging Students and Teacher Stress.
-
- nr.2 Wolf, J.C. van der & J.M.F. Touw (2005).
 Onderzoek naar zorg in het curriculum van de Theo Thijssen Academie.
-
- nr.3 Doorn, E.C. van (2005).
 De Gedragingenlijst voor Leraren (Tweede onderzoeksrapport).
-
- nr.4 Wolf, J.C. van der (2005).
 Probleemouders en de school: een onderontwikkeld terrein.
-
- nr.5 Enthoven, M. (2005).
 The contribution of the school environment to youths' resilience: A Dutch middle-adolescent perspective.
-
- nr.6 Enthoven, M.; A.C. Bouwer; J.C. Van der Wolf & A. van Peet (2005).
 Recognizing Resilience: Development and Validation of an Instrument to Recognize Resilience in Dutch Middle-Adolescents.
-
- nr.7 Velderman, H & H.A. Everaert (2005).
 Time-out or switch? (Paper presented at the ECER conference On 9 September 2005, University College Dublin).
-
- nr.8 Touw, J.M.F., J.T.E. van Beukering & H.A. Everaert (2005).
 Teachers' Personal Constructs on Problem Behaviour (Paper presented at the annual meeting of the European Educational Research Association (EERA), Dublin, Ireland, September 7-10).
-
- nr.9 Doorn, E.C. van (2005).
 Levend leren: daar ga ik voor!
-
- nr.10 J.T.E. van Beukering, J.M.F. Touw & H. Everaert (2005).
 Teachers' personal constructs on problem behaviour: towards professional development
 & Kos, P. (2005). Personal constructs on (problem) pupils: a teacher's view.
2 Papers presented at the International Practitioner Research Conference & Collaborative Action Research Network Conference (PRAR 2005), Utrecht, The Netherlands, November 4-6, 2005.
-
- nr. 11 Everaert, H. & A. van Peet (2006). Kwalitatief en kwantitatief onderzoek.
-
- Nr. 12 Everaert, H.A. & J.C. van der Wolf (2006).
 A Comparison of Stress and Burnout between Dutch General and Special Education Teachers. Paper presented at the annual meeting of the American Educational Research Association (AERA), San Francisco, USA, April 7-11, 2006.

Nr. 13 Everaert, H.A. & J.C. van der Wolf (2006). Gender Perceptions of Challenging Student Behavior and Teacher Stress.

Nr. 14 Peet, A.A.J. van (2006). Schaalconstructie.

Nr. 15 Dellevoet, S.M.E., Beukering, J.T.E. van, Everaert, H.A. & Touw, J.M.F. (2006). Evaluatie van de methode Under Construction op Instituut Theo Thijssen.

Nr. 16 Peet, A.A.J. van (2006). Q-sort. Een rangordening.

Nr. 17 Everaert, H.A. (2007). Measuring challenging student behavior. An overview of methodological properties and decisions.

Nr. 18 Everaert, H.A. & Peet, A.A.J. van (2007). Eenvoudige statistische berekeningen met behulp van Excel.

Nr. 19 Peet, A.A.J. van (2008). Enkele explorerende analyses op de Gedragingenlijst voor leraren.

Nr. 20 Everaert, H.A. & Peet, A.A.J. van (2008). Statistische toetsen in Excel.

1	Inleiding	6
2	Het tekenen van een boxplot	7
3	Cronbach's alfa (coëfficiënt van interne consistentie)	14
3.1	De berekening van Cronbach's alfa	14
3.2	Betrouwbaarheidsanalyse en item-totaalcorrelaties	17
3.3	Het formeel uitschrijven van een betrouwbaarheidsanalyse	20
4	Enkele belangrijke kansverdelingen	24
4.1	Het histogram	24
4.2	Van een histogram naar de normale verdeling	29
4.3	De normale verdeling: spelen met linker- en rechteroverschrijdingskansen	34
4.4	Van de normale naar de standaardnormale verdeling	37
4.5	De t-verdeling	39
4.6	Functies gebaseerd op de t-verdeling	44
5	Toetsen	47
5.1	Toetsingssituatie 1: t-toets voor het gemiddelde	48
5.2	Toetsingssituatie 2: t-toets voor het verschil tussen twee gemiddelden	49
5.3	Toetsingssituatie 3: t-toets voor verschillcores	50
5.4	Toetsingssituatie 4: chikwadraattoets voor kruistabellen	52

1 Inleiding

In KG-18 hebben we verschillende grafieken laten zien. Een veel gebruikte grafische figuur in de statistiek is de zogenaamde boxplot. Weliswaar wordt deze niet vermeld bij de standaard keuzemogelijkheden onder de Wizard Grafieken, toch is het mogelijk om deze figuur met behulp van Excel te tekenen. KG-20 opent met het tekenen van een boxplot. Het tweede onderwerp is het berekenen van de Cronbach's alfa, andere woorden hiervoor zijn homogeniteitsindex of coëfficiënt alfa. Beide begrippen worden zo vaak gebruikt in onderzoeksrapporten en artikelen dat we gemeend hebben dit onderwerp in dit KG te bespreken. Bovendien geeft dit ons de mogelijkheid om de veelzijdigheid van Excel te demonstreren.

Deze KG-publicatie is het logische vervolg op KG-18. In het laatste hoofdstuk van KG-18 hebben we een aantal rekenschema's gemaakt voor het berekenen van kengetallen zoals het gemiddelde, de variantie en de correlatiecoëfficiënt. Natuurlijk kunt u dergelijke kengetallen ook met een functie in Excel opvragen. Werken en 'spelen' met dergelijke rekenschema's is echter van onschatbare waarde om de betekenis van kengetallen te illustreren en eigen te maken. De lezer ziet immers meteen hoe een kengetal verandert als hij of zij de ruwe gegevens wijzigt waarop dit kengetal gebaseerd is.

Dat principe is ook weer leidend in dit KG, alleen doen we dat nu met betrekking tot kansverdelingen en dan in het bijzonder de normale verdeling, de standaardnormale verdeling en de Student t-verdeling. Er is voor gekozen om steeds de linkeroverschrijdingskansen van deze verdelingen te laten zien door te verwijzen naar cellen waarin de relevante kengetallen waarop deze kansverdelingen zijn gebaseerd, zijn opgenomen. Waar u in KG-18 leerde om kengetallen uit te rekenen, worden in dit KG deze kengetallen gebruikt in rekenschema's voor het opstellen van kansverdelingen. Wederom wordt u uitgenodigd om met deze kengetallen (eindeloos) te experimenteren. Wij zijn van mening dat dit uw begrip van de genoemde kansverdelingen ten goede komt.

De genoemde kansverdelingen, en dan met name de t-verdeling, zijn van grote praktische waarde bij kleinschalig kwantitatief onderzoek in uw eigen klas. De t-verdeling beschrijft de relatie tussen populatie en steekproef. Het is het stuk gereedschap dat u nodig kan hebben om steekproefgegevens te generaliseren naar populatieniveau. In de laatste paragraaf wordt dit geïllustreerd met drie varianten van de t-toets.

In statistische softwareprogramma's kunt u met één druk op de knop de juiste statistische maten uitrekenen. Die ene druk veronderstelt wel dat u precies weet welke knop u moet hebben en wat de uitkomsten dan vervolgens betekenen. Statistische kennis is onontbeerlijk en dat geldt voor alle relevante software. Statistiek in Excel, en zeker op de manier zoals wij dat in KG-18 en KG-20 voor ogen hebben gehad, doet naast kennis van zaken ook een beroep op een actieve houding. Het opstellen van uitgebreide rekenschema's, het zelf uitschrijven van formules en het niet werken met de specifieke statistiekmodule Gegevensanalyse (zie KG-18, pagina 69), resulteert in een relatief groot aantal handelingen of toetsaanslagen. Wij zijn van mening dat die paar extra handelingen die u in Excel moet maken, u dwingen om goed na te denken over wat u precies wilt uitrekenen en wat de uitkomst dan betekent. We hopen dat deze kleine omweg zal leiden tot een dieper begrip van de materie. Behalve door het bestuderen van theorie, rijpt uw inzicht ook door actief aan de knoppen te zitten en heel veel te oefenen met getallen.

2 Het tekenen van een boxplot

We beginnen met het tekenen van boxplots. Boxplots zijn handig voor het weergeven van diverse kengetallen in één figuur. In een boxplot worden zowel centrum- als spreidingsmaten weergegeven. Bovendien kunt u al deze kengetallen voor verschillende groepen tegelijkertijd weergeven. Stel, we hebben drie series toetsgegevens van een klas van twaalf leerlingen. We gaan van deze gegevens per toets achtereenvolgens de volgende vijf kengetallen berekenen en tekenen:

- 1^e kwartiel;
- laagste score;
- 2^e kwartiel (mediaan);
- de hoogste score;
- en het 3^e kwartiel.

Voor de drie kwartielen maken we gebruik van de functie '=KWARTIEL()'. De laagste en hoogste score bepalen we op met de functies '=MIN()' en '=MAX()'.

Misschien geniet de volgorde waarin de kengetallen zijn gesorteerd niet uw voorkeur. Die zienswijze delen we. Immers, de laagste score is per definitie bijna altijd lager dan het eerste kwartiel. Zoiets geldt ook met betrekking tot het eind van het rijtje: de hoogste score is minimaal gelijk, maar bijna altijd hoger dan het derde kwartiel. Voor het tekenen van een boxplot in Excel is de hierboven geschetste volgorde echter onvermijdelijk.¹

De gegevens die we in drie boxplots gaan tekenen, staan in Tabel 2.1 en zijn afkomstig uit KG-18.

	A	B	C	D
1	CEL A1	Toets 1	Toets 2	Toets 3
2	leerling 1	4	7	6
3	leerling 2	6	8	6
4	leerling 3	7	7	6
5	leerling 4	8	8	8
6	leerling 5	6	8	8
7	leerling 6	5	8	7
8	leerling 7	6	6	5
9	leerling 8	7	7	6
10	leerling 9	6	7	9
11	leerling 10	6	6	9
12	leerling 11	5	8	7
13	leerling 12	9	8	10

Tabel 2.1 Drie serie toetsgegevens van twaalf leerlingen.

¹ Ere wie ere toekomt: Neville Hunt op <http://www.mis.coventry.ac.uk/~nhunt/boxplot.htm>. Prachtige site met enorme veel Excel tips. Behalve de site van Hunt is ook de volgende site de moeite waard: <http://www.wiswijzer.nl/pagina.asp?nummer=233> meer dan aanbevelenswaardig, en niet alleen voor de verschillende figuren.

- Neem alle gegevens over in de cellen A1 tot en met D13.
- Typ in de cellen E2 tot en met E7 onder elkaar de volgende tekstjes 'Kengetallen', 'eerste kwartiel', 'minimum', 'tweede kwartiel (mediaan)', 'maximum' en 'derde kwartiel'.
- Typ in cel F2 'Toets 1', in cel G2 'Toets 2' en in cel H2 'Toets 3'.
- Zet een dikke streep onder de cellen E2 tot en met H2, zoals in Tabel 2.2.

E	F	G	H
	CEL E1		
Kengetallen	Toets 1	Toets 2	Toets 3
eerste kwartiel			
minimum			
tweede kwartiel (mediaan)			
maximum			
derde kwartiel			

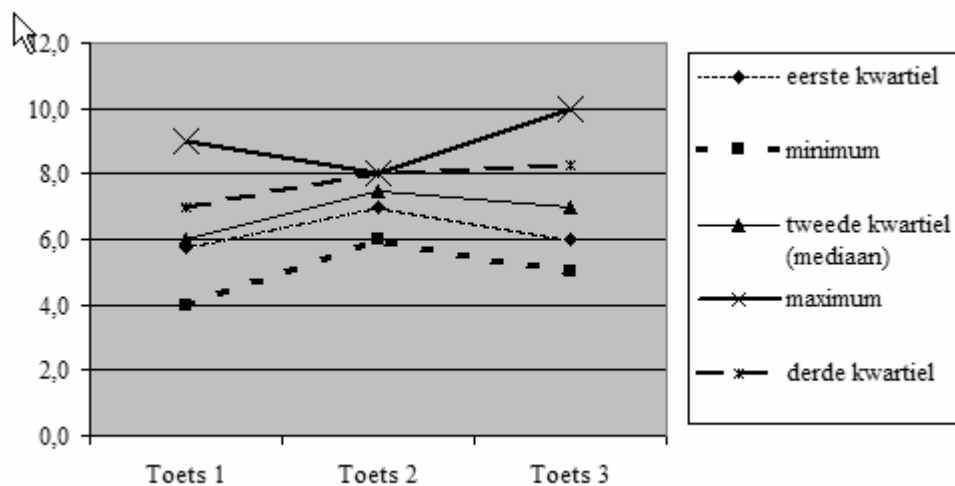
Tabel 2.2. Schema voor het berekenen van kengetallen ten behoeve van een boxplot.

- Ga met cursor in cel F3 staan en typ '=KWARTIEL(B2:B13;1)'. Deze functie kent twee argumenten. De cellen B2:B13 bevatten de getallen waarop de functie van toepassing is en getal '1' geeft aan dat we het eerste kwartiel willen uitrekenen.
- Kopieer deze functie naar de cellen G3 en H3.
- Schrijf deze functie ook in de cellen F5, G5 en H5, met dat verschil dat we niet het eerste, maar het tweede kwartiel willen. Dus in cel F5 staat '=KWARTIEL(B2:B13;2)' = 6,0.
- Herhaal dit voor het derde kwartiel op de zevende rij. In cel H7 staat dus '=KWARTIEL(D2:D13;3)'.
- Typ in cel F4 '=MIN(B2:B13)', en kopieer dit naar de cellen G4 en H4.
- Herhaal dit voor de zevende rij met de functie '=Max(B2:B13)'.
- Centreer alle getallen in de kolommen en rond af op een decimaal nauwkeurig. Het resultaat ziet u in Tabel 2.3.

E	F	G	H
	CEL E1		
Kengetallen	Toets 1	Toets 2	Toets 3
eerste kwartiel	5,8	7,0	6,0
minimum	4,0	6,0	5,0
tweede kwartiel (mediaan)	6,0	7,5	7,0
maximum	9,0	8,0	10,0
derde kwartiel	7,0	8,0	8,3

Tabel 2.3. Volgorde van relevante kengetallen voor het tekenen van drie boxplots.

- Selecteer de gehele tabel, dus van cel E2 tot en met H7: dus 'E2:H7'.
- Schakel de Wizard Grafieken aan.
- Selecteer lijnen, eerste optie en de radiobutton Rijen.
- Klik op Voltooien.
- U krijgt de onderstaande Figuur. Vermoedelijk zijn bij het uitwerken van deze oefening de markeringen en lijnen op uw scherm anders gekleurd. Met behulp van Gegevensreeks opmaken onder de rechtermuisknop kunt u dit wijzigen (zie onder).

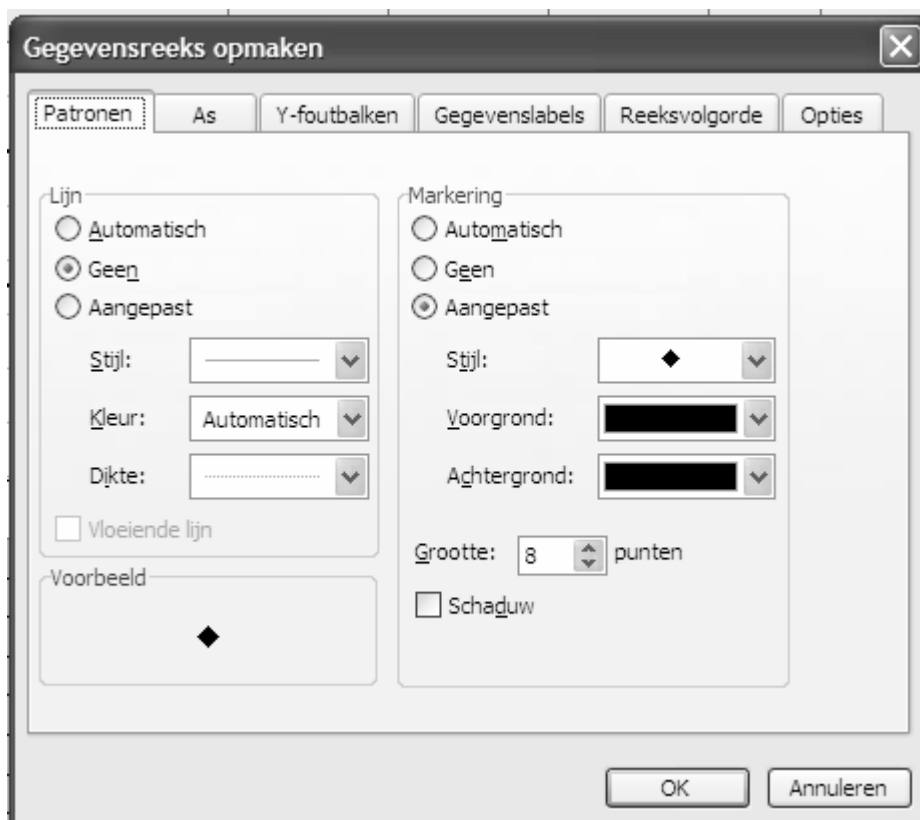


Figuur 2.4. Hard op weg naar mooie boxplots.

Figuur 2.4 geeft aan hoe een bepaald kengetal zich ontwikkelt over drie verschillende toetsen. We zien bijvoorbeeld dat het minimum van toets 1 lager is dan het minimum op toets 2 dat op zijn beurt weer lager is dan het minimum van toets 3 (Het is de enige stijgende lijn in Figuur 2.4). Het punt is dat we hier niet zozeer willen zien hoe een kengetal zich over verschillende toetsen ontwikkelt, maar dat we de relatie tussen de verschillende getallen die bij dezelfde toets horen, willen laten zien. Dat is de essentie van een boxplot. We willen dus de verbinding tussen het 1e kwartiel, de laagste score, de mediaan, het derde kwartiel en de hoogste score per toets laten zien. De lijnen in Figuur 2.4 suggereren juist een verbinding tussen de toetsen en we gaan deze lijnen dan ook verwijderen en vervangen door verticale lijnen.

- Selecteer de bovenste lijn; dus de lijn die de maxima van de drie toetsen met elkaar verbindt.
- Open nu met de rechtermuisknop door het dialoogschermb Gegevensreeks opmaken te selecteren.
- Onder Lijn kiest u de radiobutton Geen.

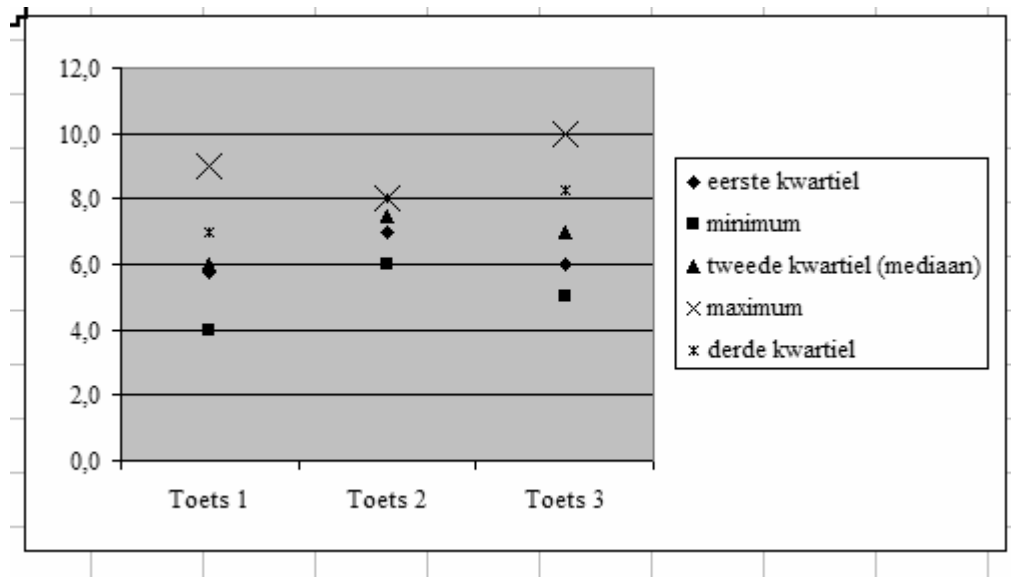
- Onder Markering selecteert u de radiobutton Aangepast; bij Stijl kiest u een ruit of salmiakje. Bij Voorggrond Zwart en voor Achtergrond Zwart. U kunt het salmiakje opblazen door het aantal punten onder Grootte op te voeren.² Zie Figuur 2.5.



Figuur 2.5. Dialoogschermb Gegevensreeks opmaken, tabblad Patronen.

- Sluit het dialoogschermb en herhaal de zojuist uitgevoerde handelingen voor de overige vier lijnen.
- Het resultaat is ongeveer zoals in Figuur 2.6.

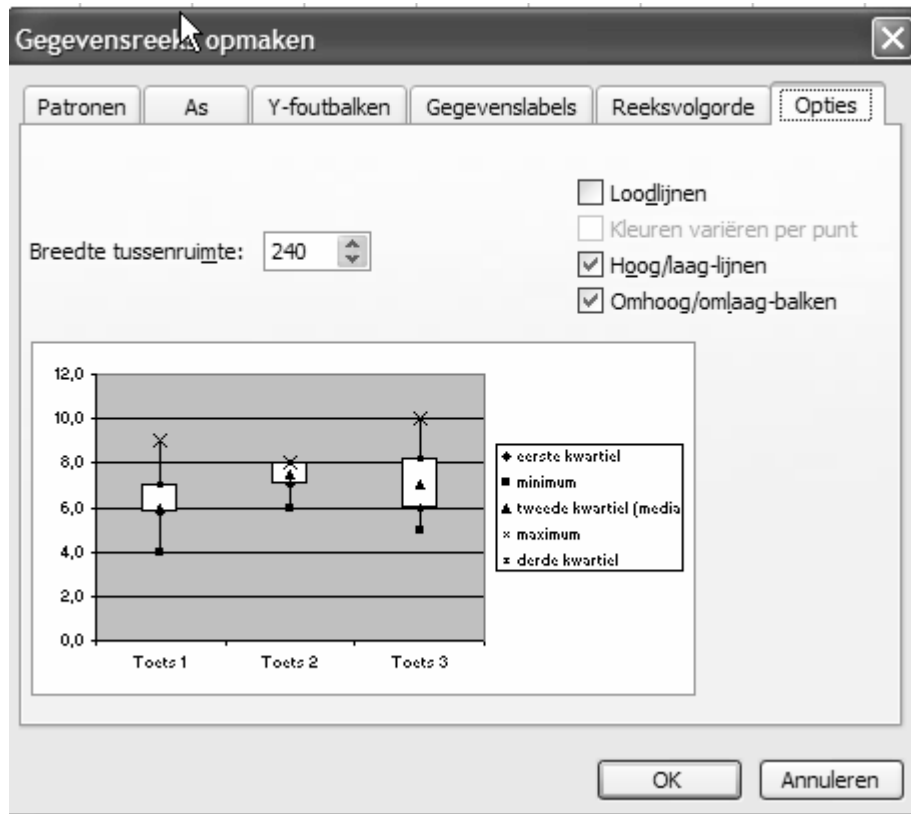
² Bij markering mag beslist 'Geen' niet gekozen worden. U heeft dan namelijk geen gegevens meer voor het Maximum in de grafiek staan.



Figuur 2.6. Bijna

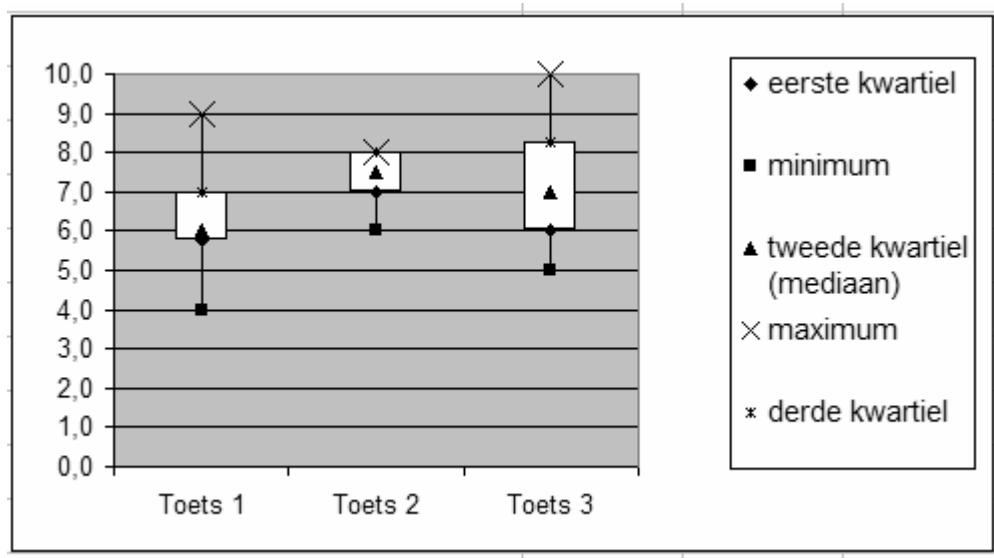
We zijn nu aardig op weg in Figuur 2.6. Nu moeten alleen nog de kengetallen per toets met elkaar verbonden worden. Dus waar we net 'min of meer horizontale' lijnen hadden, moeten nu perfect verticale lijnen worden getekend.

- We klikken op een gegevenspunt in de grafiek (maakt niet uit welke) en met de rechtermuisknop selecteren we weer het tabblad Gegevensreeks opmaken.
- Kies tabblad Opties.
- Vink Hoog/laag-lijnen aan.
- Vink ook Omhoog/omlaag-balken aan.
- Selecteer Breedte tussenruimte naar smaak: wij hebben 240 genomen.



Figuur 2.7. Dialoogscherm Gegevensreeks opmaken, tabblad Opties.

- Sluit het dialoog scherm af met een klik op de OK-knop.
- Pas het lettertype aan (Arial) in achtereenvolgens de legenda en op de beide assen.
- Een proefwerkcijfer van 12 komt u niet vaak tegen in de onderwijspraktijk. Dus pas ook de maximale waarde van de verticale as (Y-as) aan. Wijzig 12 in 10.
- Tot slot, maak eventueel gebruikte tekens voor de kengetallen groter en gebruik net als wij de kleur zwart.



Figuur 2.8. Boxplots van toetsresultaten van twaalf leerlingen naar toets.

In Figuur 2.8 zijn door het tekenen van verticale lijnen de gegevens per toets met elkaar verbonden. En volledigheidshalve, het echte werk begint nu pas: interpreteer de figuur!

3 Cronbach's alfa (coëfficiënt van interne consistentie)

In paragraaf 10.5 van het boek 'Klassenonderzoek? Dat doe je zó!' wordt de 'interne consistentie' omschreven als de mate waarin afzonderlijke items van een toets met elkaar samenhangen. Deze coëfficiënt staat bekend als Cronbach's alfa en is een schatting van de betrouwbaarheid. Coëfficiënt alfa varieert altijd tussen 0 en 1. Een hoge alfa (dus dicht bij 1) geeft aan dat een lage (hoge) score op één item gepaard gaat met lage (hoge) scores op de overige items.. Een hoge alfa zegt dus eigenlijk dat de 'gemiddelde' correlatie tussen alle items waaruit de toets is samengesteld best wel hoog is. Laten we dit maar eens gaan concretiseren met wat getallen!


3.1 De berekening van Cronbach's alfa

De formule voor het berekenen van Cronbach's alfa luidt als volgt:

$$\alpha = \frac{k}{k-1} \left[1 - \frac{\sum V_{item}}{V_X} \right]$$

waarin k aantal items
 V_{item} variantie van een item
 V_X variantie van de testscores

Als voorbeeld nemen we een wiskundetoets die bestaat uit 4 opgaven en voor elke opgave kan men maximaal 10 punten krijgen. De toets wordt afgenomen aan 8 leerlingen. De resultaten zijn door de docent in Excel gezet met de naam Tabel 3.1.

	A	B	C	D	E
1	 Cel A1	wiskundetoets			
2		a	b	c	d
3	leerling 1	4	5	7	8
4	leerling 2	6	6	8	7
5	leerling 3	6	7	6	7
6	leerling 4	8	8	9	7
7	leerling 5	10	8	7	8
8	leerling 6	6	9	8	7
9	leerling 7	4	6	7	7
10	leerling 8	9	7	8	9

Tabel 3.1. Wiskundetoetsscores van acht leerlingen per item.

Het berekenen van de itemvarianties van de afzonderlijke items kunnen we in Excel opvragen met behulp van een formule: '=VARP()'

- De variantie van de score van item a berekenen we in cel B11 met behulp van '=VARP(B3:B10)'. Het resultaat is 4,23.
- Kopieer deze cel en plak dit in de cellen C11, D11 en E11.
- Het aantal decimalen maken we in cellen B11 tot en met E11 gelijk aan 2.
- Vervolgens berekenen we in cel H12 de som van alle itemvarianties: '=SOM(B11:E11)'.

De tweede stap in de formule bestaat uit variantie van de toetsscores. De toetsscore van een leerling is gelijk aan de som van zijn of haar scores op de vier items, dus voor leerling 1 is de toetsscore: 4 + 5 + 7 + 8 = 24. Eerst stellen we vast wat de toetsscores van iedere leerling zijn. Daarna berekenen we in kolom F ook de variantie van deze toetsscore over de twaalf leerlingen die het proefwerk gemaakt hebben.

- Type in cel F2 de tekst 'Toetsscore X'. Maak de cel netjes op.
- Bereken toetsscore voor leerling 1 in cel F3: '=SOM(B3:E3)'.
- Kopieer de inhoud van cel F3 naar de overige cellen F4 tot en met F10.
- Kopieer de inhoud van cel E11 naar F11: '=VARP(F3:F10)'.
- Type in cel A14 de tekst 'Cronbach's alfa'.
- Aangezien $\alpha = \frac{4}{4-1} \left[1 - \frac{4,24 + 1,50 + 0,75 + 0,50}{12,98} \right] = 0,62$, typen we in cel C14 de formule: '=4/(4-1)*(1-SOM(B11:E11)/F11)'. Verifieer de juistheid van de formule!

	A	B	C	D	E	F	G
1	Cel A1	wiskundetoets					
2		a	b	c	d	Toetscore	
3	leerling 1	4	5	7	8	24	
4	leerling 2	6	6	8	7	27	
5	leerling 3	6	7	6	7	26	
6	leerling 4	8	8	9	7	32	
7	leerling 5	10	8	7	8	33	
8	leerling 6	6	9	8	7	30	
9	leerling 7	4	6	7	7	24	
10	leerling 8	9	7	8	9	33	
11		4,23	1,50	0,75	0,50	12,98	
12							
13							
14	Cronbach's alfa	0,62					
15							

Tabel 3.2 Schematisch overzicht van de tussenstapjes voor de berekening van Cronbach's alfa.

3.2 Betrouwbaarheidsanalyse en item-totaalcorrelaties

In deze paragraaf wordt de vraag beantwoord hoe we kunnen nagaan of de verschillende items die gebruikt worden om een bepaald begrip te meten, met elkaar samenhangen. We doen dit allereerst door de items handmatig te verwijderen en terug te plaatsen. Daarna gaan we dit op een leuke manier met behulp van de functie '=ALS()' overzichtelijk en netjes uitschrijven. Ook laten we zien hoe men de item-totaalcorrelatie kan berekenen.

In Tabel 3.3 staan de scores van vijftien studenten op acht items uit een onderzoek naar externe locus of control. De precieze verwoording van de items is uiteraard te vinden in ons theorieboek. De items zijn gescoord op een zespuntsschaal variërend van '1=geheel oneens' tot en met '6=geheel mee eens'.

Als u de gegevens uit Tabel 3.3 overneemt, dan neemt u naast kopjes en tekst, natuurlijk alleen de getallen over in het blok B4:I18. De getallen in kolom J (totaal score) als ook rij 19 kunt u met functies, kopiëren en (referentieel) plakken gemakkelijk zelf genereren. De belangrijkste formules zijn:

- Typ in cel B19 '=VARP(B4:B18)'. En kopieer dit naar overige relevante cellen in dezelfde rij.
- Typ in cel J4 '=SOM(B4:I4)'. En kopieer dit over de kolom.
- Typ in cel J19 '=VARP(J4:J18)'.
- Tot slot schrijven we de formule voor Cronbach's alfa in cel K4
'=AANTAL(B2:I2)/(AANTAL(B2:I2)-1)*(1-SOM(B19:I19)/J19)' = 0,67.

	A	B	C	D	E	F	G	H	I	J	K
1	Cel A1										
2		1	2	3	4	5	6	7	8	totaal	
3		a	b	c	d	e	f	g	h	score	Cronbach's alfa
4	persoon 1	4	4	4	4	4	6	6	1	33	0,67
5	persoon 2	2	3	1	2	2	4	2	1	17	
6	persoon 3	3	2	1	1	2	5	1	4	19	
7	persoon 4	3	3	2	2	4	3	2	2	21	
8	persoon 5	5	2	2	3	2	2	1	2	19	
9	persoon 6	1	1	1	1	1	3	1	3	12	
10	persoon 7	5	3	3	2	2	4	4	2	25	
11	persoon 8	3	4	4	3	4	4	2	2	26	
12	persoon 9	5	3	4	3	3	3	3	2	26	
13	persoon 10	4	2	2	3	3	2	2	3	21	
14	persoon 11	4	1	4	2	4	4	1	1	21	
15	persoon 12	4	2	2	3	2	4	1	2	20	
16	persoon 13	6	2	2	1	2	4	3	2	22	
17	persoon 14	3	1	3	3	3	3	1	2	19	
18	persoon 15	4	2	2	2	3	4	2	2	21	
19		1,53	0,89	1,18	0,76	0,86	1,02	1,85	0,60	21,18	

Tabel 3.3. Score van vijftien studenten op acht items die het begrip 'locus of control' pretenderen te meten.

We gaan nu kijken wat er gebeurt als we beurtelings een item weglaten. We beginnen met item A en daarna verwijderen we kolom B.

- Selecteer item a door met cursor bovenaan kolom B te gaan staan. Kolom B kleurt blauw.
- Klik in de menubalk op Bewerken gevolgd door Verwijderen.
- De waarde van Cronbach's alfa in cel K4 'blijft' ongewijzigd 0,67.
- Plaats kolom B terug door op Ongedaan maken in Werkbalk Standaard te klikken.
- Herhaal de vorige vier stappen voor item b. Het verwijderen van item b (kolom C) resulteert in een verlaging van coëfficiënt alfa van 0,67 naar 0,56.

Als u een kolom weggooit, dan verandert natuurlijk ook de formule in cel K4. We hebben immers bij het opbouwen van de formule voor alfa referentieel verwezen naar cellen. Hoewel na verwijdering van item a (kolom B) de uitkomst weliswaar hetzelfde is, is de formule wel aangepast. Oorspronkelijk hadden we in cel K4 geschreven: $=\text{AANTAL}(B2:I2)/(\text{AANTAL}(B2:I2)-1)*(1-\text{SOM}(B19:I19)/J19)=0,67$. De cellen in de formule hebben betrekkingen op alle acht items.

Na het verwijderen van kolom B wordt de formule in cel K4 automatisch gewijzigd in: $=\text{AANTAL}(B2:H2)/(\text{AANTAL}(B2:H2)-1)*(1-\text{SOM}(B19:H19)/I19)=0,67$. De cellen hebben betrekking op de items b, c, d, e, f, g en h; zeven items in plaats van acht). Dat de uitkomst onveranderd blijft, is toeval. En na het terugplaatsen van kolom B en daarna verwijderen van kolom C staat er het volgende: $=\text{AANTAL}(B2:H2)/(\text{AANTAL}(B2:H2)-1)*(1-\text{SOM}(B19:H19)/I19)=0,59$. De cellen verwijzen naar de items a, c, d, e, f, g en h; wederom zeven items met dat verschil dat nu item a weer meedoet in plaats van b.

Voor het gemak hebben we hieronder in een klein tabelletje de waarde van Cronbach's alfa weergegeven na het beurtelings verwijderen van de verschillende items.

Cronbach's alfa na verwijderen van ...							
item a	item b	item c	item d	item e	item f	item g	item h
0,67	0,59	0,57	0,62	0,62	0,66	0,54	0,78

Tabel 3.4. Cronbach's alfa na verwijderen van de diverse items.

De conclusie van Tabel 3.4 is dat item h slecht correleert met de overige zeven items. Als dit item niet meedoet, dan neemt de waarde van alfa toe van 0,67 naar 0,78.

Dat item h geen goed item is om externe locus of control te meten, kunnen we ook op een andere manier vaststellen en wel met behulp van de item-totaalcorrelatie.³ Het is een mondvoll, maar het betekent precies wat er staat. De item-totaalcorrelatie is de maat die aangeeft hoe één item correleert met de somscore van alle items. Dus de item-totaalcorrelatie van item a is de correlatie tussen de waarden van de vijftien personen op a (kolom B) en hun somscore op alle items (kolom J). Het mag ons in dit stadium geen moeite meer kosten de berekening hiervan in Excel op te vragen.

³ Strikt genomen kan men op basis van deze uitkomsten ook van mening zijn dat item h wel de locus of control meet, maar dat de overige items juist iets anders meten.

- Typ in cel A20 de tekst: 'item-totaalcorrelatie'.
- Type in cel B20 '=CORRELATIE(B4:B18;\$J4:\$J18). Let op gebruik van \$-teken. Enkel de kolom is voorzien van \$-teken.
- Kopieer cel B20 en kopieer dit naar cellen C20 tot en met J20. In cel J20 staan de correlaties van een kolom met zichzelf. Het resultaat is dus 1.
- Maak deze cellen grijs in verband met het overzicht.
- Vergelijk de scores in de cellen A20 (0,53) tot en met H20 (-0,40) in Tabel 3.5.


	A	B	C	D	E	F	G	H	I	J	K
1	Cel A1										
2		1	2	3	4	5	6	7	8	totaal	
3		a	b	c	d	e	f	g	h	score	Cronbach's alfa
4	persoon 1	4	4	4	4	4	6	6	1	33	0,67
5	persoon 2	2	3	1	2	2	4	2	1	17	
6	persoon 3	3	2	1	1	2	5	1	4	19	
7	persoon 4	3	3	2	2	4	3	2	2	21	
8	persoon 5	5	2	2	3	2	2	1	2	19	
9	persoon 6	1	1	1	1	1	3	1	3	12	
10	persoon 7	5	3	3	2	2	4	4	2	25	
11	persoon 8	3	4	4	3	4	4	2	2	26	
12	persoon 9	5	3	4	3	3	3	3	2	26	
13	persoon 10	4	2	2	3	3	2	2	3	21	
14	persoon 11	4	1	4	2	4	4	1	1	21	
15	persoon 12	4	2	2	3	2	4	1	2	20	
16	persoon 13	6	2	2	1	2	4	3	2	22	
17	persoon 14	3	1	3	3	3	3	1	2	19	
18	persoon 15	4	2	2	2	3	4	2	2	21	
19		1,53	0,89	1,18	0,76	0,86	1,02	1,85	0,60	21,18	
20	item-totaalcorrelatie	0,53	0,73	0,77	0,61	0,64	0,49	0,82	-0,40	1,00	

Tabel 3.5. Uitkomsten op de item-totaalcorrelatie in rij 20.

Cel H20 bevat een negatieve correlatiecoëfficiënt (-0,40). Dat betekent dus dat als de somscore van alle items hoger wordt, de waarde van item h juist kleiner wordt. Of, als de somscore kleiner wordt, dan wordt de score op item h groter. Item h meet dus iets heel anders dan de overige items.

- Verwijder item h.

Het resultaat staat in Tabel 3.6.

	A	B	C	D	E	F	G	H	I	J
1	Cel A1									
2		1	2	3	4	5	6	7	totaal score	Cronbach's alfa
3		a	b	c	d	e	f	g		
4	persoon 1	4	4	4	4	4	6	6	32	0,78 
5	persoon 2	2	3	1	2	2	4	2	16	
6	persoon 3	3	2	1	1	2	5	1	15	
7	persoon 4	3	3	2	2	4	3	2	19	
8	persoon 5	5	2	2	3	2	2	1	17	
9	persoon 6	1	1	1	1	1	3	1	9	
10	persoon 7	5	3	3	2	2	4	4	23	
11	persoon 8	3	4	4	3	4	4	2	24	
12	persoon 9	5	3	4	3	3	3	3	24	
13	persoon 10	4	2	2	3	3	2	2	18	
14	persoon 11	4	1	4	2	4	4	1	20	
15	persoon 12	4	2	2	3	2	4	1	18	
16	persoon 13	6	2	2	1	2	4	3	20	
17	persoon 14	3	1	3	3	3	3	1	17	
18	persoon 15	4	2	2	2	3	4	2	19	
19		1,53	0,89	1,18	0,76	0,86	1,02	1,85	24,64	
20	item-totaalcorrelatie	0,52	0,73	0,79	0,63	0,66	0,49	0,82	1,00	

Tabel 3.6 Cronbach's alfa en item-totaalcorrelatie na verwijderen van item h.

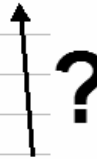
Merk op dat Cronbach's' alfa is toegenomen tot 0,78, ook de totaalscores in kolom I zijn gewijzigd. Kortom, we begonnen met een vragenlijst bestaande uit acht items en een coëfficiënt alfa van 0,67 en we eindigen dus met een vragenlijst die bestaat uit zeven items en een alfa van 0,78.

3.3 Het formeel uitschrijven van een betrouwbaarheidsanalyse

Het is niet handig om door het verwijderen en weer terug plaatsen van kolommen de gewenste Cronbach's alfa's te berekenen. Dat leidt in de praktijk te gemakkelijk tot fouten. We gaan dat in de volgende oefening niet alleen netter, maar vooral ook systematischer opschrijven. Dit gebeurt door een aantal cellen te definiëren waarin aangegeven wordt of een kolom wel (waarde=1) of niet (waarde=0) moet worden meegenomen in de berekening. We maken hiervoor gebruik van de functie '=ALS()', uiteraard in combinatie met handig absoluut en referentieel verwijzen (zoals altijd dus ☺).

- Kopieer de cellen A1 tot en met K19 in Tabel 3.6.
- Plak al deze cellen in blok L1:V19.
- Vervang in cel L1 de tekst 'Cel A1' door de tekst 'Cel L1'.
- Verifieer de juistheid van de formule voor het bereken van coëfficiënt alfa in cel V4: '=AANTAL(M2:T2)/(AANTAL(M2:T2)-1)*(1-SOM(M19:T19)/U19)'. De uitkomst is 0,67, maar dat wist u al.

- Ga naar cel L2 en typ daarin de tekst 'Staat dit item wel of niet aangeschakeld?'.
- Selecteer de cellen L2 en L3 samen. Vul de tekst uit door achtereenvolgens Opmaak, celeigenschappen en Uitlijning te selecteren. Kies onder horizontaal: Uitvullen.
- Vervang de getallenreeks (1,2,3,4,5,6,7 en 8) in cel M2 tot en met T2 door acht enen (dus 1,1,1,1,1,1,1 en 1).
- Vervang in cel M4 de het getal 4 door de formule =ALS(\$M\$2=1;B4;0). Dit moet u als volgt lezen. Als de waarde van cel M2 gelijk is aan 1, dan wil ik het getal in cel B4 terug. In alle andere gevallen, dus als de waarde van cel B2 ongelijk aan 1 is, geef dan het getal 0 terug.
- Let op: in cel M4 wordt absoluut verwezen naar cel M2. U kopieert cel M4 en plakt dit in de cellen N4 tot en met T4.
- Vervolgens zorgt u ervoor dat er niet steeds naar M2, maar ook naar N2, O2, et cetera verwezen wordt. U gaat met de cursor in cel N4 staan en vervangt =ALS(\$M\$2=1;C4;0) door =ALS(\$N\$2=1;C4;0). De referentiële verwijzing naar B4 is automatisch vervangen door C4. Herhaal dit voor cellen O4 tot en met T4 en in cel T4 komt dus uiteindelijk '=ALS(\$T\$2=1;I4;0)' te staan.
- Kopieer de cellen M4 tot en met T4.
- Plak dit over de cellen M5 tot en met T18. U ziet ook meteen waarom we zojuist gekozen hebben om de cellen op de tweede rij te bevriezen.
- Er is visueel nog helemaal niets veranderd op uw scherm. Ook alfa is nog steeds 0,67 in cel V4.
- Vervang in cel M2 het getal 1 door 0. Alle cellen in kolom M veranderen in 0.
- Let op: ook de waarden in kolom U zijn veranderd.
- Vervang in de cellen M2 tot en met R2 steeds het getal 1 in 0. Het resultaat ziet u in Tabel 3.7

L	M	N	O	P	Q	R	S	T	U	V
Cell L1	items								totaal score	Cronbach's alfa
Staat dit item wel of niet aangeschakeld?	a	b	c	d	e	f	g	h		
persoon 1	0	0	0	0	0	0	6	1	7	<div style="border: 1px solid black; padding: 5px; text-align: center;">-0,57</div> <div style="text-align: center; margin-top: 10px;">  </div>
persoon 2	0	0	0	0	0	0	2	1	3	
persoon 3	0	0	0	0	0	0	1	4	5	
persoon 4	0	0	0	0	0	0	2	2	4	
persoon 5	0	0	0	0	0	0	1	2	3	
persoon 6	0	0	0	0	0	0	1	3	4	
persoon 7	0	0	0	0	0	0	4	2	6	
persoon 8	0	0	0	0	0	0	2	2	4	
persoon 9	0	0	0	0	0	0	3	2	5	
persoon 10	0	0	0	0	0	0	2	3	5	
persoon 11	0	0	0	0	0	0	1	1	2	
persoon 12	0	0	0	0	0	0	1	2	3	
persoon 13	0	0	0	0	0	0	3	2	5	
persoon 14	0	0	0	0	0	0	1	2	3	
persoon 15	0	0	0	0	0	0	2	2	4	
	0,00	0,00	0,00	0,00	0,00	0,00	1,85	0,60	1,63	

Tabel 3.7. Cronbach's alfa is berekend - en nog fout ook- op basis van twee items.

In de figuur ziet u zes kolommen (M tot en met R) met enkel het getal 0 in iedere cel. Misschien dat het lijkt dat in cel U4 de som van de cellen S4 en T4 staat, maar dat is niet helemaal juist. In cel U4 staat de som van de cellen M4 + N4 + + S4 + T4 = 0 + 0 + ... + 6 + 1 = 7. De waarde van Cronbach's alfa in cel V4 blijkt negatief te zijn (-0,57). Dat kan helemaal niet; aan het begin van dit hoofdstuk hebben we gezegd dat alfa varieert van 0 tot 1. Er staat dus nog ergens een fout in de formule voor Cronbach's alfa in cel V4.

- Ga naar cel V4 en verander de formule
'=AANTAL(M2:T2)/(AANTAL(M2:T2)-1)*(1-SOM(M19:T19)/U19)'
in '=SOM(M2:T2)/(SOM(M2:T2)-1)*(1-SOM(M19:T19)/U19)'.

De functie '=AANTAL()' geeft namelijk het aantal cellen waarin een getal staat. We willen niet langer het *aantal* cellen waarin een getal staat meenemen in de formule voor alfa, maar we willen het aantal items dat echt meedoet bij de bepaling van alfa. Door gebruik te maken van de cijfers 1 en 0 *sommeert* de functie =SOM() nu het juiste aantal items. Een soortgelijk trucje hebben we ook gezien in KG-18.

- Verifieer de juistheid van de formule in cel V4 door in de tweede rij enkel celwaarden 1 te plaatsen. Alfa verandert dan weer in 0,67. Dat is geen nieuws!
- Vervang tot slot in cel T2 het getal 1 door 0 (item h doet dus niet langer mee). Cronbach's alfa verandert in 0,78 (En ook dat wist u al).

U kunt nu dus allerlei items een voor een of zelfs meerdere tegelijkertijd uitschakelen. Is het mogelijk om alfa nog verder te verhogen door behalve item h ook nog een ander item te verwijderen? Zo ja, welk item dan?

4 Enkele belangrijke kansverdelingen

In deze paragraaf beginnen we met het tekenen van een histogram op basis van vierhonderd waarnemingen. Vervolgens wordt ‘dezelfde tekening’ ook getekend met behulp van twee kengetallen (het steekproefgemiddelde en de steekproefstandaarddeviatie) die gebaseerd zijn op deze vierhonderd waarnemingen. Deze tweede tekening is beter bekend onder de naam van de normale verdeling. We gaan deze normale verdeling nader onderzoeken door steeds linker- en rechteroverschrijdingskansen te berekenen onder variërende omstandigheden. Vervolgens stappen we van de normale verdeling over op de standaard normale verdeling, ook wel z-verdeling genoemd.

Nadat we een paar eigenschappen van de normale en standaardnormale verdeling hebben laten zien, richten we onze aandacht op de relatie tussen de populatie en een aselechte steekproef uit deze populatie. Een aselechte steekproef wordt getrokken om het populatiegemiddelde te schatten. Het steekproefgemiddelde fungeert als schatter voor het populatiegemiddelde en we staan stil bij de relatie tussen beide gemiddelden.⁴ Deze relatie kunnen we beschrijven met behulp van een andere kansverdeling: de t-verdeling. De grap is nu dat als we een hele grote aselechte steekproef zouden trekken, de t-verdeling rond het steekproefgemiddelde weer overgaat in een z-verdeling.

De t-verdeling is praktisch van groot belang bij het uitvoeren van kleinschalig (statistisch) onderzoek in uw eigen klas of school. We ronden dit hoofdstuk af met een overzicht van drie Excel-functies gebaseerd op deze t-verdeling. Deze functies hebben we nodig om in het laatste hoofdstuk te kunnen ‘toetsen’.

4.1 Het histogram

Stel we weten van vierhonderd vrouwen de leeftijd waarop ze een kind hebben gekregen. Deze gegevens staan in frequentietabel 4.1.

⁴ De relatie tussen de standaarddeviatie in de populatie en steekproef wordt niet behandeld in dit KG.

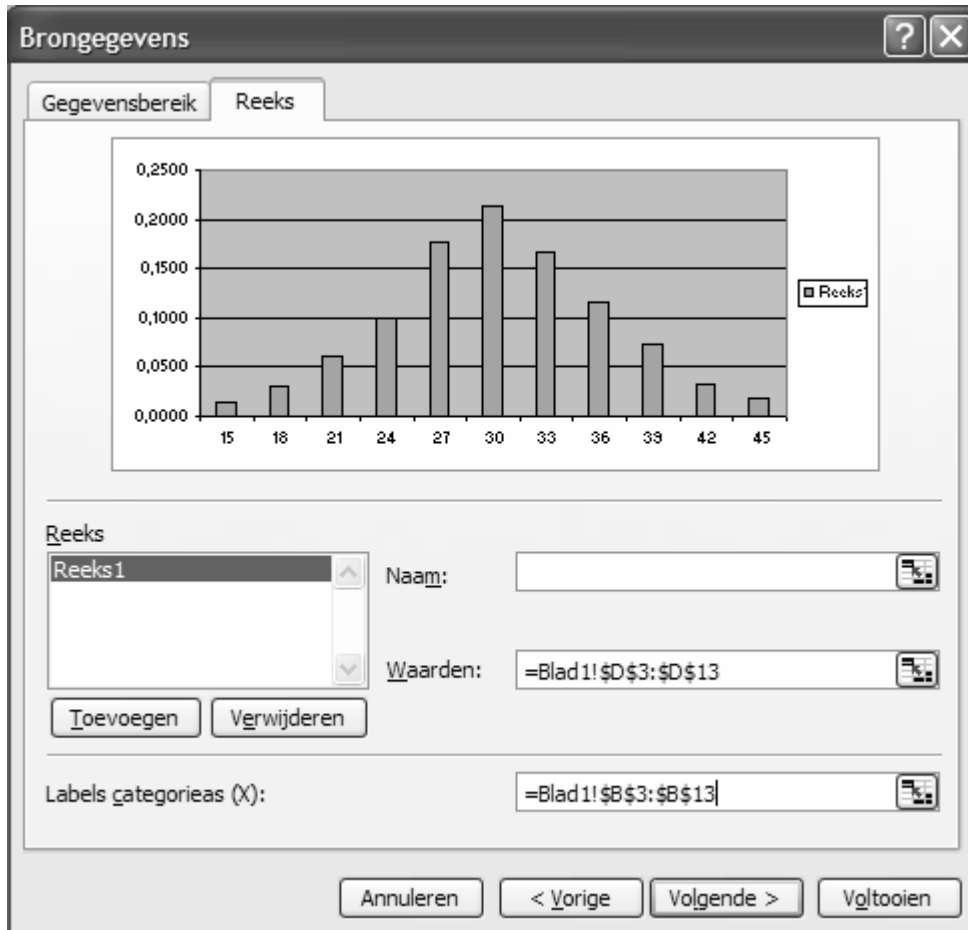
	A	B	C	D	E
1	Cel A1				
2	Leeftijdklassen	Klassenmidden	f	Proportie	Cumulatieve proportie
3	14-16	15	6	0,0150	0,0150
4	17-19	18	12	0,0300	0,0450
5	20-22	21	24	0,0600	0,1050
6	23-25	24	40	0,1000	0,2050
7	26-28	27	71	0,1775	0,3825
8	29-31	30	85	0,2125	0,5950
9	32-34	33	67	0,1675	0,7625
10	35-37	36	46	0,1150	0,8775
11	38-40	39	29	0,0725	0,9500
12	41-43	42	13	0,0325	0,9825
13	44-46	45	7	0,0175	1,0000
14			400	1	

Tabel 4.1. Frequentietabel van de leeftijd waarop vierhonderd vrouwen een kind hebben gekregen.

- Neem Tabel 4.1 over in Excel. Het spreekt inmiddels voor zich dat u de gegevens in de kolommen D (Proportie) en E (Cumulatieve proportie) niet overtypt, maar zelf berekent op basis van gegevens in kolom C.

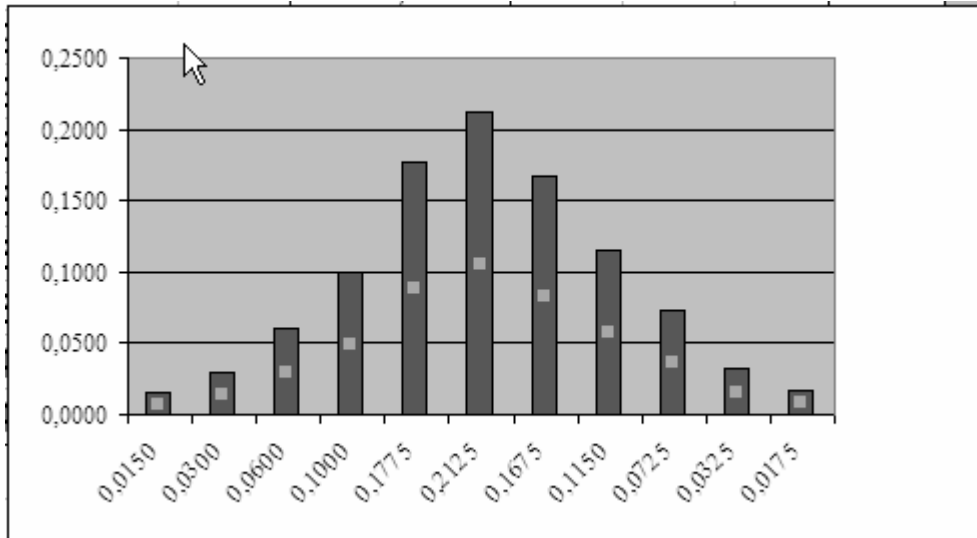
Een histogram is in wezen niets anders dan een staafdiagram waarvan de staven tegen elkaar geplakt zijn. Het is de grafische weergave van een gegroepeerde frequentieverdeling. Behalve de frequentie van de variabele kan men er ook voor kiezen om de proporties weer te geven. We hebben in dit KG voor proporties gekozen.

- Open de Wizard Grafieken en selecteer Grafiektype Kolom.
- Geef in het tabblad Gegevensbereik van dialoogscherm Brongegevens de cellen D3 tot en met D13: '=Blad1!\$D\$3:\$D\$13'.
- Klik in Brongegevens op tabblad Reeks.
- Ga met cursor naar Labels categorieas (X): en selecteer met de muis de cellen B3 tot en met B13. Zie Figuur 4.2.



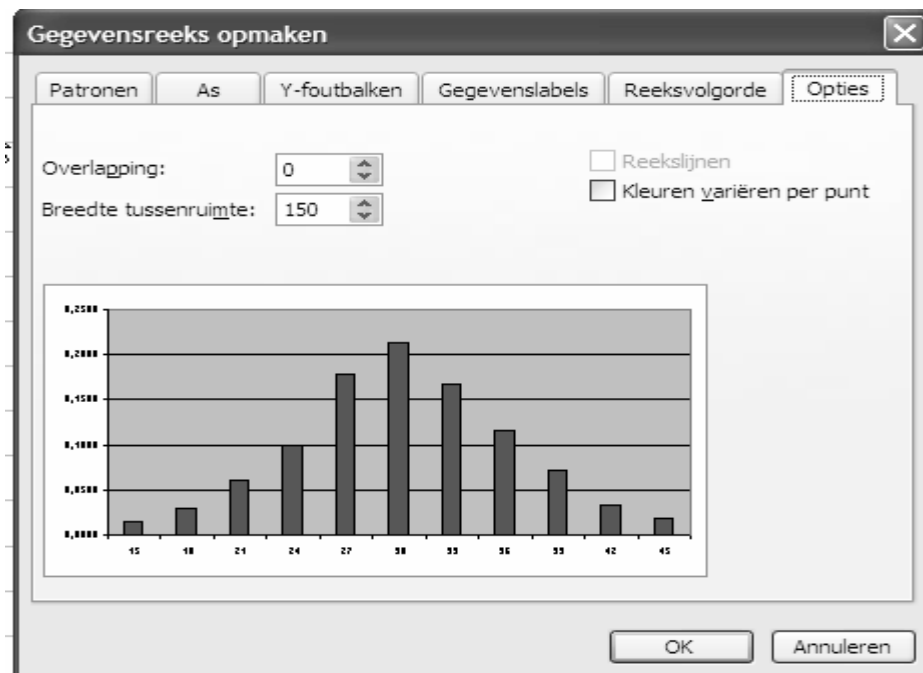
Figuur 4.2. Tabblad Reeks van dialoogscherm Brongegevens na selecteren van de juiste cellen in kolom D bij Labels categorieas (X):

- Sluit Figuur 4.2 door op Voltooien te klikken. De gegevens worden nu in een staafdiagram weergegeven.
- Verwijder de legenda met daarin Reeks1 en selecteer een van de staafjes door erop te klikken. In alle staafjes verschijnt een klein vierkantje. Zie Figuur 4.3.



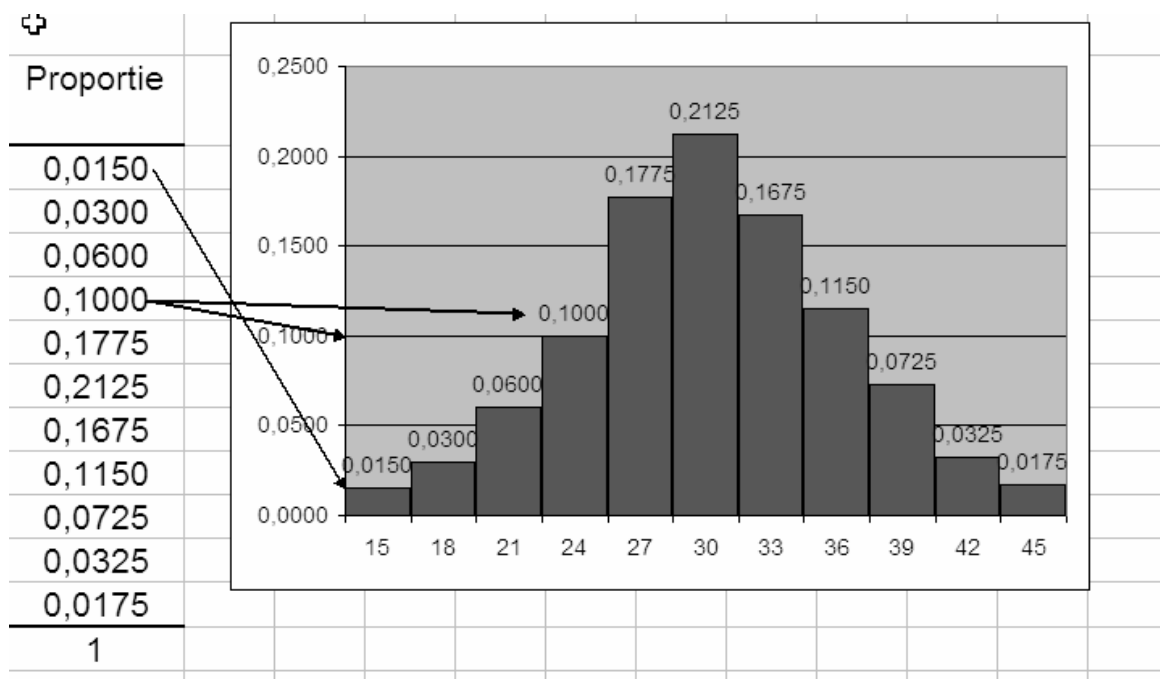
Figuur 4.3. Van staafdiagram naar histogram.

- Dubbelklik op een klein vierkantje. Het dialoogschermd Gegevensreeks opmaken wordt geopend.
- Selecteer het tabblad Opties (zie Figuur 4.4).
-



Figuur 4.4. Tabblad Opties van dialoogschermd Gegevensreeks opmaken.

- Verander Breedte tussenruimte: (in Figuur 4.4 is deze 150) in 0.
- Selecteer nu het tabblad Gegevenslabels.
- Vink Waarde af en sluit het tabblad door op OK te klikken.
- Van elk staafje is de waarde uit kolom E zichtbaar. Zie Figuur 4.5.



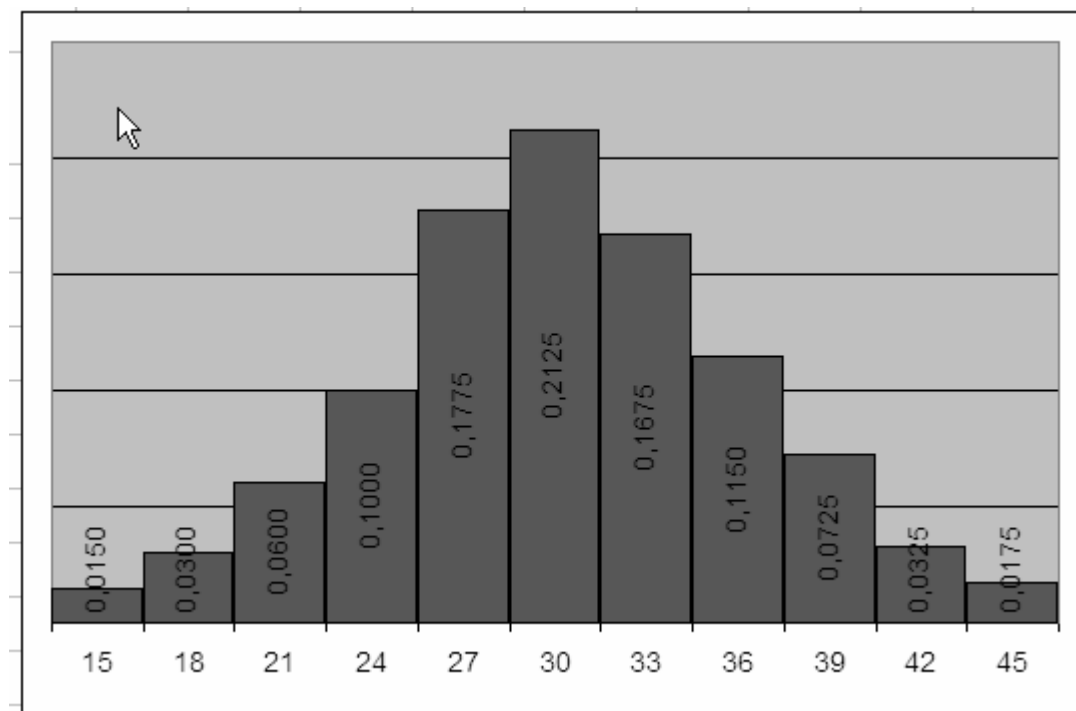
Figuur 4.5. Histogram van de leeftijd waarop vierhonderd vrouwen een kind hebben gekregen (1).

Figuur 4.5 mag nu officieel en met recht een histogram genoemd worden en u weet dat de totale paarse oppervlakte precies gelijk aan 1 is. En dat weet u omdat de som van alle proporties (zie cel D14) namelijk gelijk aan 1 is. Er is echter iets 'optisch bedriegelijks' in Figuur 4.5.

In Figuur 4.5 wordt namelijk de hoogte van de staafjes aangegeven op de verticale as (Y-as). We hebben dit willen verduidelijken voor twee staafjes, namelijk het staafje met klassenmidden 15 jaar en het staafje met klassenmidden 24 jaar. Die laatste is het duidelijkst want 'de hoogte' is precies 0,1000. Het punt is echter dat we doordat we proporties hebben getekend, eigenlijk met een oppervlaktemaat werken (totale oppervlakte is 1, zie hierboven). De oppervlakte van het staafje met klassenmidden 24 is 0,1000 deel van het geheel. Om deze precieze betekenis van het histogram gebaseerd op proporties beter voor het voetlicht te krijgen, passen we een cosmetische foefje toe. We verwijderen de Y-as en zetten de oppervlaktecijfers midden in de staafjes.

- Ga naar de grafiek en klik met de rechtermuisknop precies op de Y-as.
- Er verschijnt een heel klein schermje en daarvan klikt u op 'Wissen', waardoor de Y-as verdwijnt.
- Dubbelklik op een Waarde in de grafiek.
- Selecteer het tabblad Uitlijning van dialoogschermb Gegevenslabels opmaken.
- Bij Labelpositie kiest u de optie 'Centreren'.

- In het rechterdeel van dit dialoogschermb, staat een rechthoek met daarin een halve cirkel. Met uw cursor pakt u het lijntje in deze cirkel en dit schuift u 90 graden omhoog. Het resultaat ziet u hieronder in Figuur 4.6.



Figuur 4.6. Histogram van de leeftijd waarop vierhonderd moeders een kind hebben gekregen (2).

Misschien dat u van mening bent dat er niets veranderd is ten opzichte van de vorige figuren. Dat klopt, maar wel is veel duidelijker zichtbaar dat de oppervlakte van de staafjes in plaats van de hoogte de kern van de zaak is.⁵

4.2 Van een histogram naar de normale verdeling

We gaan nu op zoek naar een manier om Figuur 4.6 op een andere manier te tekenen. Dat wil zeggen, we zoeken een mooie lijn die min of meer door de topjes van de staven loopt en waarvan de oppervlakte begrensd door deze lijn en de as waarop de leeftijden staan vermeld (X-as), ook gelijk aan 1 is. Deze lijn bestaat en wordt bepaald door twee kengetallen gebaseerd op de frequentieverdeling: het gemiddelde en de standaarddeviatie. Om te kijken of de door ons gevonden lijn ook echt past worden de cumulatieve proporties van de frequentieverdeling in Tabel 4.1/Figuur 4.6 vergeleken met de linkeroverschrijdingskansen die bij deze lijn horen. Dit hele verhaal gaan we hieronder uitvoeren in drie afzonderlijke stappen:

⁵ Op zich is de hoogte hier niet fout. Alle hoogtes bij elkaar opgeteld zijn namelijk ook gelijk aan 1. Dat is in ons voorbeeld zo omdat alle staafjes even breed zijn.

Stap 1. Het is bekend dat de lijn die we zoeken getekend kan worden met een ingewikkelde formule en dat we daar maar twee kengetallen (in plaats van vierhonderd) voor nodig blijken hebben: het gemiddelde en de standaarddeviatie. Die kengetallen gaan we berekenen op basis van Tabel 4.1.

Stap 2. We tekenen deze lijn op basis van het onder stap 1 gevonden gemiddelde en standaarddeviatie. Aangezien het tekenen van deze lijn vrij lastig is, presenteren we dit plaatje zonder verder commentaar.

Stap 3. Een ingewikkeld plaatje opnemen is niet de beste manier om de lezer ervan te overtuigen dat de door ons getekende lijn erg goed bij Tabel 4.1/Figuur 4.6 past. We willen u overtuigen door steeds de kans te berekenen dat een moeder op bepaalde leeftijd een kind heeft gekregen op basis van zowel Tabel 4.1 als ook op basis van onze 'mooie' theoretische lijn. Deze kansen worden steeds met elkaar vergeleken.

Stap 1. Het berekenen van het gemiddelde en standaarddeviatie is in het huidige voorbeeld iets gecompliceerder dan in KG-18 is beschreven. Dat komt omdat in KG-18 de gegevens steeds per individu zijn weggeschreven. In Tabel 4.1 zijn de gegevens per leeftijdsklasse weggeschreven en die klassen zijn niet allemaal even groot. Er zijn nu immers zes moeders in de leeftijdsklasse met midden 15; twaalf moeders in de leeftijdsklasse met midden 18, etc. We moeten dus het aantal moeders per klasse meewegen bij het berekenen van het gemiddelde en standaarddeviatie. Gelukkig bestaat hiervoor een handige functie in Excel: '=SOMPRODUCT()'.

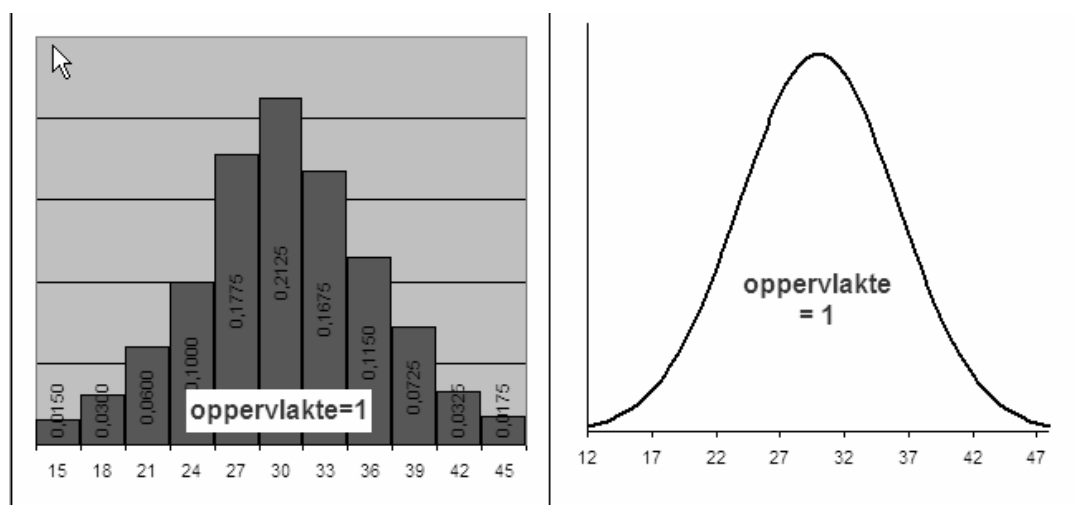
- Typ in cel A17 de tekst 'Gemiddelde leeftijd'.
- Typ in cel B17 '=SOMPRODUCT(B3:B13;C3:C13)'. Deze functie geeft het resultaat van de cellen $B3 \cdot C3 + B4 \cdot C4 + B5 \cdot C5 + \dots + B13 \cdot C13 = 15 \cdot 6 + 18 \cdot 12 + 21 \cdot 24 + \dots + 45 \cdot 7 = 12096$. Zo houden we rekening met de verschillende aantallen moeders per leeftijdsklasse. Het *gewogen* gemiddelde vinden we vervolgens door 12096 te delen door 400 = 30,24 (cel C17).
- Ook voor het berekenen van standaarddeviatie gebruiken we de functie '=SOMPRODUCT()'. We zouden in cel F2 graag de tekst 'klassenmidden minus gemiddelde en dat dan in het kwadraat' typen. Dat is namelijk in woorden hoe u begint met het berekenen van de standaarddeviatie. Dat kan korter, typ in cel F2: ' $(x - \text{gem}(x))^2$ '.
- Bereken in cel F3 het verschil tussen klassenmidden in B3 en het gemiddelde in cel C17 en vermenigvuldig dit verschil met zichzelf. Dus in cel F3 typt u ' $=(B3-\$C\$17)*(B3-\$C\$17)$ '.
- Kopieer deze formule ' $=(B3-\$C\$17)*(B3-\$C\$17)$ ' naar de cellen F4 tot en met F13.
- Typ in cel A18 de tekst 'Standaarddeviatie'.
- Maak in cel B18 gebruik van de functie SOMPRODUCT() door de cellen C3 tot en met C13 te vermenigvuldigen met F3 tot en met F13. Dus '=SOMPRODUCT(C3:C13;F3:F13)'. Het resultaat moet zijn 15132,96.
- Deel in cel C18 door het aantal waarnemingen min 1 en trek hieruit de wortel: ' $=\text{WORTEL}(B18/(400-1))$ '. De standaarddeviatie is 6,16. Uiteraard ronden we alles weer netjes af!

Stap 1 is nu klaar. De vierhonderd moeders hebben gemiddeld op hun dertigste een kind gekregen met een standaarddeviatie van iets meer dan zes jaar. Het resultaat van de gehele oefening staat in Tabel 4.7.

	A	B	C	D	E	F
1	Cell A1					
2	Leeftijdsklassen	Klassenmidden	f	Proportie	Cumulatieve proportie	$(x - \text{gem}(x))^2$
3	14-16	15	6	0,0150	0,0150	232,26
4	17-19	18	12	0,0300	0,0450	149,82
5	20-22	21	24	0,0600	0,1050	85,38
6	23-25	24	40	0,1000	0,2050	38,94
7	26-28	27	71	0,1775	0,3825	10,50
8	29-31	30	85	0,2125	0,5950	0,06
9	32-34	33	67	0,1675	0,7625	7,62
10	35-37	36	46	0,1150	0,8775	33,18
11	38-40	39	29	0,0725	0,9500	76,74
12	41-43	42	13	0,0325	0,9825	138,30
13	44-46	45	7	0,0175	1,0000	217,86
14			400	1		
15						
16	Gemiddelde leeftijd	12096	30,24			
17	Standaarddeviatie	15132,96	6,16			

Tabel 4.7. Rekenschema voor vaststellen van het *gewogen* gemiddelde en standaarddeviatie met behulp van functie '=SOMPRODUCT()'.

Stap 2. Met behulp van de twee kengetallen berekend in Stap 1 (30,24 en 6,16) is onze 'mooie' lijn getekend in het rechterdeel van Figuur 4.8.⁶



⁶ Alle normale en standaardnormale verdeling in KG-20 zijn getekend met behulp van commando's geschreven door Jacob van Peet, mei 2001.

Figuur 4.8. Links. De steekproefgegevens getekend op basis van vierhonderd waarnemingen. Rechts de kansverdeling gebaseerd op de berekende kengetallen (gemiddelde=30,24 en standaarddeviatie=6,16) van de steekproef.

De vraag hoe goed beide plaatjes in Figuur 4.8 bij elkaar passen, gaan we in stap 3 beantwoorden.

Stap 3. De vraag hoe goed de beide figuren bij elkaar passen, beantwoorden we door de oppervlakten van beide figuren met elkaar te vergelijken. We weten dat oppervlakte in beide plaatjes gelijk is aan 1. Maar hoe is die oppervlakte precies verdeeld? We gaan nu steeds de beide oppervlakten links van een bepaalde leeftijdsgrens met elkaar vergelijken. Die oppervlakte aan de linkerkant van een bepaalde waarde (leeftijd) noemen we de linkeroverschrijdingskans. Een linkeroverschrijdingskans is bijvoorbeeld de kans dat een moeder van 16 jaar oud of jonger een kind heeft, of de kans dat een moeder van 19 jaar oud of jonger een kind heeft, of de kans dat een moeder jonger dan 34 jaar een kind heeft, of de kans dat een moeder jonger dan 60 jaar een kind heeft. Formeel kunnen we de zojuist gegeven voorbeelden als volgt opschrijven: $P(x \leq 16)$, $P(x \leq 19)$, $P(x \leq 34)$ en $P(x \leq 60)$. Op basis van de gegevens in Tabel 4.1 kunt u de gevraagde linkeroverschrijdingskansen direct aflezen uit kolom E. Ga zelf na dat: $P(x \leq 16) = 0,0150$; $P(x \leq 19) = 0,0450$; $P(x \leq 34) = 0,07802$ en $P(x \leq 60) = 1$.

We hoeven dus alleen nog maar te bepalen wat de linkeroverschrijdingskansen zijn in het rechterplaatje van Figuur 4.8. Daarvoor bestaat een functie: '=NORM.VERD()'. Deze functie kent vier argumenten, waarvan u er al met twee kennis heeft gemaakt: het gemiddelde en de standaarddeviatie. Het derde en vierde argument hebben betrekking op de gevraagde oppervlakte aan de linkerkant. U moet dus de waarde invoeren (bijvoorbeeld 16 jaar) en de vraag of alle 'denkbeeldige' staafjes links van die waarde bij elkaar moeten worden opgeteld (=1) of niet (=0). Het getal 1 gebruiken we om de cumulatieve kansverdeling willen berekenen. Het getal 0 wordt gebruikt als we de kans op precies één bepaalde leeftijd willen berekenen. Dit vierde en laatste argument (1 of 0) geeft het verschil aan tussen $P(x \leq 16)$ respectievelijk $P(x = 16)$. Als u nu wilt weten wat de kans is dat een moeder van 16 jaar oud of jonger een kind heeft gekregen, dan typt u '=NORM.VERD(16;30,24;6,16;1)'.⁷ De volgorde van de argumenten voor de functie NORM.VERD() is als volgt: eerst de waarde van x, dan het gemiddelde gevolgd door de standaarddeviatie en tot slot het getal 0 of 1).

- Ga naar Tabel 4.7 en kopieer de waarden in de cellen C16 en C17. Selecteer Bewerken in de menubalk, kies Plakken speciaal en kies radiobutton bij Waarden gevolg door een klik op OK. U hebt hiermee uw referentiële verwijzingen en functies in cellen C16 en C17 door twee vaste getallen vervangen.
- Aangezien u kolom F niet meer nodig heeft voor het berekenen van kengetallen, kunt u deze kolom verwijderen om in de 'nieuwe' kolom F gaan werken. Typ in cel F2 de tekst 'Cumulatieve theoretische kans'. Zet ook een dikke streep onder deze cel.
- Ga met de cursor in cel F3 staan en typ '=NORM.VERD(B3+1;\$C\$16;\$C\$17;1)'. Het eerste argument is de waarde van cel B3 (klassenmiddelen) plus een jaar extra (=B3+1). We doen dit omdat de uiterste grens van deze klasse geen 15 (=cel B3), maar 16 jaar (=cel B3+1) is. De functie heeft vervolgens het gemiddelde (C16) en de standaarddeviatie (C17) nodig. Als laatste argument gebruiken we het cijfer 1 om aan te geven dat we de cumulatieve kans willen. In cel F3 verschijnt het de uitkomst 0,0104.

⁷ Functieargumenten in Excel voor het bereken van de functie NORM.VERD(X; μ ; σ ; 0 of 1).

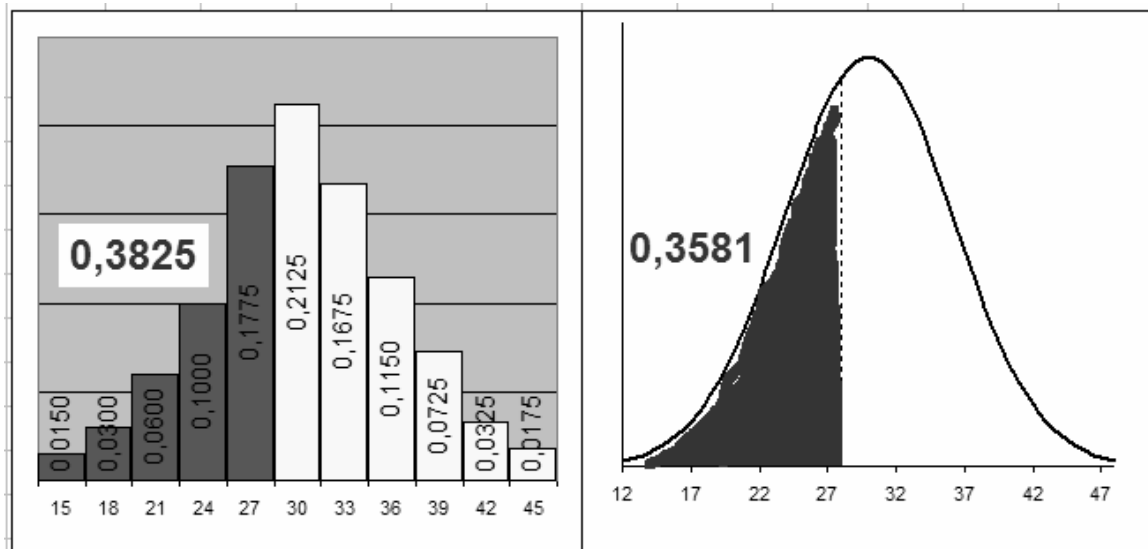
- Kopieer cel F3 en plak dit over de cellen F4 tot en met F13. In cel F13 staat de functie '=NORM.VERD(B13+2;\$C\$16;\$C\$17;1)' = 0,9967 en dat is nagenoeg gelijk aan 1, zoals we verwacht hadden.

In Tabel 4.9 hebben we alle gegevens naast elkaar gezet.

	A	B	C	D	E	F
1	Cel A1					
	Leeftijdklassen	Klassenmidden		Proportie	Cumulatieve proportie	Cumulatieve theoretische kans
2			f			
3	14-16	15	6	0,0150	0,0150	0,0104
4	17-19	18	12	0,0300	0,0450	0,0340
5	20-22	21	24	0,0600	0,1050	0,0905
6	23-25	24	40	0,1000	0,2050	0,1975
7	26-28	27	71	0,1775	0,3825	0,3581
8	29-31	30	85	0,2125	0,5950	0,5491
9	32-34	33	67	0,1675	0,7625	0,7292
10	35-37	36	46	0,1150	0,8775	0,8638
11	38-40	39	29	0,0725	0,9500	0,9435
12	41-43	42	13	0,0325	0,9825	0,9808
13	44-46	45	7	0,0175	1,0000	0,9947
14			400	1		
15						
16		gemiddelde	30,24			
17		standaarddeviatie	6,16			

Tabel 4.9. Vergelijking tussen de cumulatieve proportie op basis van vierhonderd waarnemingen en cumulatieve theoretische kansverdeling op basis van twee afgeleide kengetallen.

Tot slot, moeten we nog beoordelen hoe goed de theoretische verdeling op basis van de twee kengetallen past bij de vierhonderd waarnemingen waarop de twee kengetallen zijn afgeleid. U kunt dit zien door de waarden in de kolommen E en F met elkaar te vergelijken. In Figuur 4.10 hebben we de betekenis van de cellen E7 en F7 nog eens getekend. Als u moeite heeft met deze Figuur raden we u aan om het gehele verhaal nog eens na te lezen en elk cijfer dat u in Figuur 4.10 tegenkomt voor uzelf te verklaren.



Figuur 4.10. De kans dat een moeder een kind kreeg voor haar 28^{ste} levensjaar. Rechts geven we dit aan met de cumulatieve proportie. Links is dit gedaan door de linkeroverschrijdingskans te bepalen.

De conclusie is dat op basis van oppervlakte (=kansverdeling!) de beide tekeningen in Figuur 4.10 goed bij elkaar passen.⁸

4.3 De normale verdeling: spelen met linker- en rechteroverschrijdingskansen

Met behulp van de functie '=NORM.VERD()' konden we, voor welke leeftijd van de moeders dan ook, steeds de linkeroverschrijdingskansen bepalen. Ook de rechteroverschrijdingskansen kan eenvoudig worden uitgerekend. Dat is namelijk: $1 - \text{de linkeroverschrijdingskans}$. Dit laatste impliceert dat de kans op precies één waarneming nagenoeg 0 zal zijn.⁹

Om wat meer inzicht in het verloop van normale verdeling te krijgen gaan we hieronder een schema in opstellen waarmee we naar hartenlust het gemiddelde en de standaarddeviatie kunnen veranderen. Stel, we weten dat jongens in de populatie gemiddeld 180 cm lang zijn, en dat de standaarddeviatie 20 cm is. Hoe groot is nu de kans dat we één jongen in de collegebanken aantreffen die kleiner is dan 160 cm. En wat gebeurt er met die kans als het gemiddelde niet 180 maar 175 blijkt te zijn? Of als de standaarddeviatie geen 20, maar 10 cm of 25 cm is.

- Typ in cel B1 de tekst 'Lengte jongens in de populatie'.
- Typ in cel A2 de Griekse letter mu ($=\mu$). De letter μ gebruiken we om het populatiegemiddelde van een variabele aan te geven.
- In cel B2 typen we de gemiddelde lengte in centimeters van de populatie jongens in Nederland. Typ: 180 in cel B2.

⁸ Er bestaat ook een formele toets om na te gaan of een frequentie verdeling ook echt normaal verdeeld is. Die laten we in dit KG buiten beschouwing.

⁹ Reken maar na voor Tabel 4.9 '=NORM.VERD(B3+1;\$C\$16;\$C\$17;0)'. Het laatste argument geeft aan dat we precies de kans willen weten op de waarde in cel B3 +1, dus $P(x=16)$.

- Typ in cel A3 de Griekse letter sigma ($=\sigma$). De letter σ gebruiken we om de standaarddeviatie van een variabele aan te geven.
- In cel C2 plaatsen we de standaarddeviatie van de populatie jongens in Nederland. Typ in cel B2 20. Let op: als de lengte in centimeters is uitgedrukt, dan is het logisch om de standaarddeviatie eveneens in centimeters uit te drukken.
- Typ in cel C1 een hoofdletter 'X'. Deze X symboliseert een bepaalde waarde van de variabele in de steekproef. We trekken een steekproef met lengte 1, namelijk één jongen uit de Nederlandse populatie jongens.
- Typ in cel D1 de tekst 'linkeroverschrijdingskans'. We willen in kolom D bepalen hoe groot de kans is dat we een jongen tegenkomen kleiner dan lengte X, als we weten dat het populatiegemiddelde μ (en standaarddeviatie σ) is.
- Typ in cel E1 de tekst 'rechteroverschrijdingskans'. We willen weten hoe groot de kans is dat we een jongen tegenkomen langer dan lengte X, als we weten dat het populatiegemiddelde μ (en standaarddeviatie σ) is.
- Typ in cel C2 het getal '130'.
- Ga met cursor naar cel C3 en typ '=C2+5'; In cel C3 staat dus $130 + 5 = 135$. Zo meteen gaan we cel C3 ook kopiëren en plakken over de kolom C. Bedenk welke waarde u terugkrijgt als u C3 kopieert en plakt naar C4? En vervolgens naar C5?
- In cel D2 gaan we de kans berekenen dat deze jongen in de steekproef kleiner dan of gelijk is aan 130 cm als verder geldt dat het populatiegemiddelde $\mu = 180$ en de standaarddeviatie $\sigma = 20$. U zou dus kunnen typen '=NORM.VERD(130;180;20;1)'. Maar deze kengetallen van de populatie staan ook in de cellen B2 en B3. We typen dus in cel D2 '=NORM.VERD(C2;\$B\$2;\$B\$3;1)'. Let op gebruik dollartekens.
- Typ in cel E2 '=1-D2', dat is de rechteroverschrijdingskans bij het zelfde gemiddelde en standaarddeviatie.
- Kopieer de cellen D2 en E2 en plak beide cellen over D3 en E3.
- Kopieer nu de cellen C3 tot en met E3 en plak dit over de cellen C4 tot en E16. Het resultaat staat afgebeeld in Tabel 4.11.

	A	B	C	D	E
1		lengte van jongens in de populatie	\bar{X}	linker overschrijdingskans	rechter overschrijdingskans
2	μ	180	130	0,01	0,99
3	σ	20	135	0,01	0,99
4			140	0,02	0,98
5			145	0,04	0,96
6			150	0,07	0,93
7			155	0,11	0,89
8			160	0,16	0,84
9			165	0,23	0,77
10			170	0,31	0,69
11			175	0,40	0,60
12			180	0,50	0,50
13			185	0,60	0,40
14			190	0,69	0,31
15			195	0,77	0,23
16			200	0,84	0,16

Tabel 4.11. De normale verdeling met het populatiegemiddelde $\mu=180$ en standaarddeviatie $\sigma=20$.

Hoe groot is kans dat een jongen kleiner dan of precies 155 centimeter is? Antwoord: 11%.

Hoe groot is kans dat een jongen kleiner dan of precies 180 centimeter is? Antwoord: 50%.

Hoe groot is kans dat een jongen groter is dan of precies 175 centimeter is? Antwoord: 60%.

Hoe groot is de kans dat een jongen groter is dan of precies 200 centimeter? Antwoord: 16%.

Vervolgens gaat u eindeloos variëren met het populatiegemiddelde. Het idee er achter is steeds hetzelfde: kijk goed wat er gebeurt met de kansverdeling.

- Typ achtereenvolgens in cel B2: 160 175, 190, 200 en 230 (dat is wel heel lang).
- Beantwoord steeds de bovenstaande vier vragen. Dus bij een populatiegemiddelde van 230 is de kans dat we iemand ontmoeten die langer is dan of gelijk aan 200 centimeter maar liefst 97%. Logisch, want als de gemiddelde lengte in de populatie veel langer wordt, dan wordt de kans om een slungel te ontmoeten natuurlijk ook groter.

- Aangezien onze afstanden in kolom C niet reëel zijn bij een populatiegemiddelde van 230 cm, vervangen we in cel C2 het getal 130 door 180.
- Ga terug naar de uitgangpositie in Tabel 4.11 (Cel 'B2=180'; cel B3 '=20' en cel C3 '=130').
- Nu gaan we de standaarddeviatie veranderen. Typ achtereenvolgens de getallen 10 en 5 in en beantwoord weer dezelfde vragen wat er gebeurt met de kans een kleine of grote jongen tegen te komen. Hieronder staan de antwoorden bij Cel 'B2=180'; cel B3 '=10' en cel C3 '=130'
 - o Hoe groot is kans dat een jongen kleiner dan of precies 155 centimeter is? Antwoord: 1%.
 - o Hoe groot is kans dat een jongen kleiner dan of precies 180 centimeter is? Antwoord: 50%.
 - o Hoe groot is kans dat een jongen groter is dan of precies 175 centimeter is? Antwoord: 69%.
 - o Hoe groot is de kans dat een jongen groter is dan of precies 200 centimeter? Antwoord: 2%.
- Vervolgens maken we de standaarddeviaties in de populatie weer groter. Dat wil zeggen de onderlinge verschillen tussen de jongens in de populatie groter worden: er is meer variëteit. Hieronder staan de antwoorden bij Cel 'B2=180'; cel B3 '=40' en cel C3 '=130'
 - o Hoe groot is kans dat een jongen kleiner dan of precies 155 centimeter is? Antwoord: 27%.
 - o Hoe groot is kans dat een jongen kleiner dan of precies 180 centimeter is? Antwoord: 50%. (Nog steeds!)
 - o Hoe groot is kans dat een jongen groter is dan of precies 175 centimeter is? Antwoord: 55%.
 - o Hoe groot is de kans dat een jongen groter is dan of precies 200 centimeter? Antwoord: 31%.

4.4 Van de normale naar de standaardnormale verdeling

We hebben hierboven gewerkt met de normale verdeling, een andere verdeling is de standaardnormale verdeling. De standaardnormale verdeling is hetzelfde als de normale verdeling, met dat verschil dat het gemiddelde naar 0 en de standaarddeviatie naar 1 is omgerekend. In de hoofdstukken 7 en 12 van het boek 'Klassenonderzoek? dat doe je zó!' kunt u dit nalezen. In principe kan elke normale verdeling worden herleid tot een standaardnormale verdeling door middel van een lineaire transformatie. Dit herleiden gebeurt in twee stapjes. Eerst trekken we van elke individuele waarneming het gemiddelde af en vervolgens delen we door de standaarddeviatie.

- Verwijder kolom E (de rechteroverschrijdingskansen).
- Ga naar cel E1, type 'z'. Zet een dikke streep onder deze cel.
- Ga naar cel E2, type '=(C2-\$B\$2)/\$B\$3' =-2,50. Het gebruik van dollartekens verraadt alvast dat we gaan kopiëren: plak inhoud van cel E2 over cellen E3 tot en met E16.
- In cel F1 typen we nu 'linkeroverschrijdingskansen'. Ook hier komt een dikke streep onder de cel.

- In cel F2 gaan we weer de functie voor de normale verdeling gebruiken voor het bepalen van de linkeroverschrijdskansen in kolom F. Het verschil is alleen dat we nu met een gemiddelde 0 en standaarddeviatie 1 gaan rekenen. Dus typ in cel F2 '=NORM.VERD(E2;0;1;1). Het resultaat staat in Figuur 4.12.
- Voor de duidelijkheid voegen we helemaal bovenin een extra rij in. In de nieuwe samengevoegde cellen C1 en D1 typen 'normale verdeling' en in de nieuwe samengevoegde cellen E1 en F1 typen we 'standaardnormale verdeling'. (Al kunt u dit strikt genomen ook aflezen uit het gebruik van de letters X en z).

	A	B	C	D	E	F
1			normale verdeling		standaardnormale verdeling	
2		lengte jongens in de populatie	X	linker overschrijdskans	z	linker overschrijdskans
3	μ	180	130	0,01	-2,50	0,01
4	σ	20	135	0,01	-2,25	0,01
5			140	0,02	-2,00	0,02
6			145	0,04	-1,75	0,04
7			150	0,07	-1,50	0,07
8			155	0,11	-1,25	0,11
9			160	0,16	-1,00	0,16
10			165	0,23	-0,75	0,23
11			170	0,31	-0,50	0,31
12			175	0,40	-0,25	0,40
13			180	0,50	0,00	0,50
14			185	0,60	0,25	0,60
15			190	0,69	0,50	0,69
16			195	0,77	0,75	0,77
17			200	0,84	1,00	0,84

Figuur 4.12. De normale verdeling ($\mu=180$; $\sigma=20$) wordt herleid tot de standaardnormale verdeling ($\mu=0$; $\sigma=1$).

Wat u verder ook verandert in de velden B2, B3 of C3. De waarden in de cellen in de kolommen D en F zijn altijd precies gelijk aan elkaar. Dus wat u ook verandert: cel D3=F3; of D7=F7; of D16=F16; of D10=F10. Ga dit na tot u voor uzelf, en wel hardop, onder woorden kunt brengen waarom dit het geval is. Op basis van de uitkomsten in de D en F kunnen we steeds beantwoorden hoe groot de kans dat

we een iemand tegenkomen die kleiner (linkeroverschrijdingskans) is dan een bepaalde waarde. We hebben nu twee verdelingen besproken (de normale en de standaardnormale verdeling).¹⁰

4.5 De t-verdeling

In de vorige paragraaf zijn we steeds uitgegaan van een bekende populatie. We wisten het populatiegemiddelde van de lengte van Nederlandse jongens en we kenden de populatiestandaarddeviatie. Vervolgens konden we de kans op een jongen kleiner of groter dan een bepaalde lengte uitrekenen. De praktijk is weerbarstiger: we zullen vaak een aselechte steekproef moeten trekken om het populatiegemiddelde en de populatiestandaarddeviatie te schatten. En vervolgens wilt u niet weten wat de kans is dat u één jongen met een bepaalde lengte tegen het lijf loopt, maar meerdere jongens met een bepaalde gemiddelde lengte. Dat probleem pakken we aan met behulp van de t-verdeling.

De t-verdeling (er zijn er overigens meerdere) is tot stand gekomen doordat men uit een bekende populatie steeds opnieuw steekproeven van dezelfde lengte heeft getrokken. Op grond van al deze steekproeven wordt steeds opnieuw de t-waarde berekend. Bij het berekenen van deze t-waarde wordt zowel rekening gehouden met het steekproefgemiddelde, de standaarddeviatie als ook de lengte van de steekproef. Vervolgens is in kaart gebracht hoe de frequentieverdeling van al deze steekproefgemiddelden er nu precies uit ziet. Het idee is nu dat de verdeling van deze t-waarde ons iets kan vertellen over het populatiegemiddelde.

Dit principe kunnen we in Excel illustreren door steeds opnieuw een steekproef te trekken uit een serie getallen variërend van 1 tot en met 999. De populatie bestaat dus uit 999 unieke getallen en daarvan bereken we allereerst het populatiegemiddelde en de standaarddeviatie. Vervolgens trekken we een aselechte steekproef van 50 getallen uit deze populatie, berekenen het steekproefgemiddelde en steekproefstandaarddeviatie. Daarna trekken we een nieuwe steekproef uit deze zelfde populatie, berekenen weer het steekproefgemiddelde en steekproefstandaarddeviatie, enz, enz. Uiteindelijk hebben we dan een verzameling steekproefgemiddelden waarvan we per steekproef de t-waarde vaststellen (standaardiseren!). We gaan dit idee hieronder schematisch weergeven.

- typ in cel A1 de tekst 'Populatie'.
- Typ in cel A2 het getal 1.
- Typ in cel A3 '=A2+1'.
- Kopieer cel A3 en plak dit over de cellen A4 tot en met A1000. U hebt nu een kolom getallen oplopend van 1 tot en met 999.

¹⁰ Er bestaat uiteraard ook een aparte functie voor de standaard normale verdeling '`=STAND.NORM.VERD()`'. Het argument van deze functie is de z-waarde. Dus in plaats van '`=NORM.VERD(E2;0;1;1)`' kunt u ook meteen '`=STAND.NORM.VERD(E2)`' opvragen. Het scheelt drie argumenten (namelijk gemiddelde 0, standaarddeviatie 1 en cumulatief) en het resultaat is identiek. U kunt deze functie gebruiken in plaats van een tabel met standaardnormale verdelingen. Een tweede functie waar we u in dit verband op willen wijzen is '`=STAND.NORM.INV()`'. U voert hier de linkeroverschrijdingskans als argument in en de functie retourneert de bijbehorende z-waarde. Kortom, stel $z = -1$, dan geeft '`=STAND.NORM.VERD(-1)`' u de kans 0,1587 terug en als u daarna deze kans invoert in de functie '`=STAND.NORM.INV(0,1587)`' dan krijgt u het getal -1 terug.

- Typ in cel B2 'populatiegemiddelde'
- Gebruik in cel C2 de formule '=GEMIDDELDE(A2:A1000)'. De uitkomst is 500,00. Rond af op twee decimalen.
- Typ in cel B3 'populatiestandaarddeviatie'.
- In cel C2 gebruiken we de formule '=STDEVP(A2:A1000)'. Let op de letter P in STDEVP(). De populatievariantie van 999 getallen oplopend van 1 tot en met 999 is 288,39.
- Typ in cel D1 '50 random getallen'.
- We gaan nu één random getal genereren van 1 tot en met 999 in cel D2. Typ daartoe in cel D2 '=AFRONDEN((ASELECT()*1000);0)'. De functie 'ASELECT()' genereert een random getal tussen 0 en 1. Aangezien we getallen willen die 1000 keer groter zijn, vermenigvuldigen we de uitkomst van ASELECT()*1000. Bovendien willen we gehele getallen. We ronden dus ook nog eens netjes af. Het resultaat is steeds een getal tussen de 1 en 999.¹¹
- Kopieer de cel D2 en plak dit over de cellen D3 tot en met D51. U krijgt nu 50 random nummers.
- Typ in cel E1 het woord 'trekking'.
- Typ in cel F1 de tekst 't-waarde'.
- Typ in cel E2 het getal '1' gevolgd door een harde return. U ziet de 50 getallen in kolom D veranderen.
- Typ in de cellen B5, B6 en B7 achtereenvolgens 'steekproefgemiddelde', 'steekproefstandaarddeviatie' en 't- waarde'.
- Typ in cel C5 '=GEMIDDELDE(D2:D51)', hiermee berekent u het gemiddelde over de 50 random getallen.
- Typ in cel C6 '=STDEV(D2:D51)'. Deze functie geeft u de standaarddeviatie terug van de 50 random getrokken getallen. Let op: STDEV() is *niet* het zelfde als STDEVP(). De functie STDEV() wordt gebruikt om de standaarddeviatie van de steekproef te berekenen, terwijl STDEVP() wordt toegepast bij het bepalen van de standaarddeviatie van de populatie. Bij grote steekproeven is het verschil verwaarloosbaar.
- In cel C7 berekenen we de t-waarde: $t = \frac{\bar{x} - \mu}{s_x / \sqrt{n}} = \frac{\bar{x} - \mu}{s_x} * \sqrt{n}$. Het steekproefgemiddelde staat in cel C5; het populatiegemiddelde in cel C2 en de standaarddeviatie van de steekproef is uitgerekend in cel C6. Het berekenen van de t-waarde gebeurt door middel van '=((C6-C2)/C6)*WORTEL(50)'. *Realiseer expliciet dat het principe van het standaardiseren van de t-waarde identiek is aan het principe van het standaardiseren van de normale verdeling.*
- Ga nu nog een keer met cursor naar cel E2, typ een '1', gevolgd door een harde return.
- Neem de waarde in cel C8 over in cel F2.

¹¹ We zijn ons bewust van een kleine afrondingsfout: ook de random getallen 0 en 1000 worden random gegenereerd. Dat is het gevolg van het gebruik van de functie '=AFRONDEN()' in combinatie met '=ASELECT()'. We zouden dit kunnen oplossen door in geval van 0 of 1000 gewoon een nieuw random getal te genereren.

- Typ een '2' in cel E3 en neem de waarde van cel C7 over in cel F2. Herhaal dit een aantal malen. We hebben uiteindelijk 17 trekkingen van 50 random getallen verricht en de even zovele t-waarden in kolom F overgenomen. Het resultaat staat in Tabel 4.12.

	A	B	C	D	E	F
1	<u>Populatie</u>	<u>berekeningen</u>	<u>kengetallen</u>	<u>50 random getallen</u>	<u>trekking</u>	<u>t-waarde</u>
2	1	populatiegemiddelde	500,00	202	1	0,36
3	2	populatiestandaarddeviatie	288,39	873	2	-1,37
4	3			438	3	-1,68
5	4			718	4	-1,20
6	5	steekproefgemiddelde	553,12	14	5	0,80
7	6	steekproefstandaarddeviatie	271,25	202	6	-1,52
8	7	t- waarde	1,3848	922	7	1,08
9	8			511	8	1,21
10	9			988	9	0,98
11	10			366	10	0,05
12	11			120	11	1,32
13	12			397	12	-0,50
14	13			419	13	-2,67
15	14			947	14	2,04
16	15			223	15	-1,75
17	16			133	16	0,63
18	17			91	17	1,3848
19	18			557		

Tabel 4.13. Opzetje voor het genereren van t-waarden op basis van zeventien verschillende steekproef met lengte 50 uit een populatie van 999 getallen variërend van 1 tot en met 999.

U kunt Tabel 4.13 verder net zo fraai maken als u zelf wilt. Ideaal zou zijn om de steeds wisselende t-waarde per trekking automatisch vast te leggen in een kolom en daarvan vervolgens weer de frequentieverdeling te tekenen. Deze frequentieverdeling van t-waarden is niks anders dan de t-verdeling en geeft ons inzicht in de relatie tussen het onbekende populatiegemiddelde en het berekende steekproefgemiddelde. Behalve gemiddelde en standaarddeviatie speelt ook de lengte van de steekproef een rol in deze relatie.

De t-verdeling is op te vragen met behulp van een functie in Excel, namelijk '=T.VERD()'. De functie T.VERD() kent drie argumenten. Het eerste argument is de waarde van t. Het tweede argument is het aantal vrijheidsgraden. Dat is niets anders dan de lengte van de steekproef minus 1. Dus als we 50 random getallen genereren, dan is het aantal vrijheidsgraden $50-1=49$.¹² Het derde argument is de vraag of de overschrijdingskans een- of tweezijdig moet worden uitgerekend. Dat laatste geeft u aan met gebruik van het getal 1 of het getal 2.

We willen nu de kansen op een bepaalde t-waarde gaan vergelijken met de kansen op een bepaalde z-waarde zoals we dit in Tabel 4.12 gedaan hebben. Dat is in eerste instantie lastiger dan gedacht. We hebben in Tabel 4.12, maar ook in de aanloop steeds met linkeroverschrijdingskansen gewerkt en aan dat principe houden we hier graag vast.

Moeilijkheid 1: Helaas is de linkeroverschrijdingskans voor getallen kleiner dan 0 niet rechtstreeks op te vragen met de functie T.VERD(). Dat wil zeggen, als we een getal kleiner dan 0 in voeren, geeft de functie een foutmelding. Aangezien de t-verdeling - net als de standaardnormale verdeling - perfect

¹² Het begrip 'vrijheidsgraden' wordt doorgaans afgekort met de letters 'df' van degrees of freedom.

symmetrisch is rond 0 is, weten we dat de linkeroverschrijdingskans even groot is als de rechteroverschrijdingskans. Zo is bijvoorbeeld $P(t \leq -1,25)$ even groot als $P(t \geq 1,25)$. We gaan dit eerste 'probleem' dus oplossen door de absolute waarde van getallen kleiner dan 0 te nemen. De linkeroverschrijdingskans $P(t \leq -1,25)$ is niet gedefinieerd voor de functie T.VERD(). Immers, -1.25 is kleiner dan 0. Dus rechteroverschrijdingskans voor $t \geq 1,25$ vragen we op met het commando '=T.VERD(1,25;49;1)' = 0,1086. En de linkeroverschrijdingskans voor $t \leq -1,25$ verkrijgen we door de absolute waarde van -1,25 te nemen. Alles bij elkaar: $P(t \leq -1,25) = \text{'T.VERD(ABS(-1,25);49;1)'} = 0,1086$.

Moeilijkheid 2: Een ander, tweede verschil met de (standaard) normale verdeling is dat we door middel van het laatste argument aan konden geven of we een cumulatieve linkeroverschrijdingskans wilden of niet. Deze theorie speelt uiteraard bij de t-verdeling ook een rol, maar in de functie voor de t-verdeling wordt er standaard van uit gegaan dat we altijd met cumulatieve kansen werken. U valt hier dus niets te kiezen. De functie T.VERD() geeft of de rechteroverschrijdingskans of de tweezijdige overschrijdingskans (dus de linker- en rechteroverschrijdingskans bij elkaar opgeteld). We hebben steeds met eenzijdige kansen gewerkt, dus het vierde argument van de functie T.VERD() is 1 en hiervan hebben we bij het berekenen van de kansen $P(t \leq -1,25)$ even groot als $P(t \geq 1,25)$ ook gebruik gemaakt. U moet niet denken dat een linkeroverschrijdingskans enkel bestaat voor getallen kleiner dan 0. In plaats van $P(t \leq -1,25)$ kan u ook gevraagd worden $P(t \leq +1,25)$. De linkeroverschrijdingskans van 1,25 is nu gelijk aan $P(t \leq 1,25) = 1 - P(t \geq 1,25) = 1 - 0,1086$ en die vragen we op met de functie '=1-T.VERD(1,25;49;1)' = 0,8914. Wederom zijn we op dit moment slechts geïnteresseerd in eenzijdige overschrijdingskansen.

Nu gaat u bovenstaande principes toepassen in een tabel waarin we steeds de lengte van de steekproef kunnen variëren opdat u de linkeroverschrijdingskansen volgens de standaardnormale en de t-verdeling met elkaar kunt vergelijken. Neem Tabel 4.14 over.

	A	B	C	D	E
1	Cel A1	linkeroverschrijdingskansen			
2		standaardnormale	t-verdeling	t-verdeling	t-verdeling
3		verdeling	n	n	n
4	z of t		5	15	25
5	-2,75				

Tabel 4.14. Opzet voor genereren van linkeroverschrijdingskansen.

- Typ in cel D4 '=C4 + 10'.
- Typ in cel E3 '=C3+20'. U begrijpt dat we de t-verdelingen willen laten zien bij drie verschillende steekproeflengtes (namelijk $n=5$, $n=15$ en $n=25$).
- Typ in de cel A6 '=A5+0,25'.
- Kopieer cel A6 en plak deze over de cellen A7 tot en met A27. In stapjes van 0,25 gaat u van dan -2,75 (cel A7) naar + 2,75 (cel A27).
- Typ in cel B5 '=NORM.VERD(A5;0;1;1)'.
- Kopieer cel B5 en plak dit in de cellen B6 tot en met B27. U kent deze kolom al uit Tabel 4.12.

- Typ in cel C5 '=T.VERD(ABS(\$A5);C\$4-1;1)'. Dus u vraagt de positieve waarde op van cel A5: namelijk $-(-2,75)$. De lengte van de steekproef staat in C4 en is gelijk aan 5. Het aantal vrijheidsgraden is dan $5-1=4$. Met het argument 'C\$4 -1' geeft u 4 vrijheidsgraden op. U ziet nu de dus $P(t \leq -2,75) = 0,0257$ bij 4 vrijheidsgraden.
- Kopieer C5 naar D5. Dus '=T.VERD(ABS(\$A5);D\$4-1;1)'. Dus nu geldt $P(t \leq -2,75) = 0,0078$ bij 14 vrijheidsgraden.
- Herhaal dit voor cel E5.
- Kopieer cel C5 tot en met E5 en plak dit over de cellen C6 tot en met E16. Afronden op twee decimalen is verstandig, anders ziet u wel heel veel detail maar geen grote lijnen. In ieder geval moeten alle linkeroverschrijdingskansen kleiner of gelijk aan 0,50 zijn. Nog scherper: enkel in rij 16 staat vier keer het zelfde getal: 0,50. We hebben nu het eerste trucje achter de rug. Voor het berekenen van linkeroverschrijdingskansen voor getallen kleiner of gelijk aan 0, nemen we de absolute waarde van de t-waarde '=T.VERD(ABS(t-waarde);vrijheidsgraden;1)'.
- De tweede 'truc' passen we eerst toe in cel C17: '=1-T.VERD(\$A17;C\$4-1;1)'. Dus voor het berekenen van de linkeroverschrijdingskans voor getallen groter dan 0 nemen we 1- '=T.VERD(t-waarde;vrijheidsgraden;1)'.
- Kopieer cel C17 naar D17 en E17.
- Selecteer cellen C17 tot en met E17 en kopieer dit in een keer over het blok: C18 tot en met E27. Het resultaat, vier series linkeroverschrijdingskansen, staat in Tabel 4.15. Rond af op twee decimalen.

	A	B	C	D	E
1	Cell A1	<div> <div></div> <div>linkeroverschrijdingskansen</div> </div>			
2		standaardnormale verdeling	t-verdeling	t-verdeling	t-verdeling
3			n	n	n
4	z of t		5	15	25
5	-2,75	0,00	0,03	0,01	0,01
6	-2,50	0,01	0,03	0,01	0,01
7	-2,25	0,01	0,04	0,02	0,02
8	-2,00	0,02	0,06	0,03	0,03
9	-1,75	0,04	0,08	0,05	0,05
10	-1,50	0,07	0,10	0,08	0,07
11	-1,25	0,11	0,14	0,12	0,11
12	-1,00	0,16	0,19	0,17	0,16
13	-0,75	0,23	0,25	0,23	0,23
14	-0,50	0,31	0,32	0,31	0,31
15	-0,25	0,40	0,41	0,40	0,40
16	0,00	0,50	0,50	0,50	0,50
17	0,25	0,60	0,59	0,60	0,60
18	0,50	0,69	0,68	0,69	0,69
19	0,75	0,77	0,75	0,77	0,77
20	1,00	0,84	0,81	0,83	0,84
21	1,25	0,89	0,86	0,88	0,89
22	1,50	0,93	0,90	0,92	0,93
23	1,75	0,96	0,92	0,95	0,95
24	2,00	0,98	0,94	0,97	0,97
25	2,25	0,99	0,96	0,98	0,98
26	2,50	0,99	0,97	0,99	0,99
27	2,75	1,00	0,97	0,99	0,99

Tabel 4.15. De z-verdeling vergeleken met drie t-verdelingen (df=4,14 en 24). De symmetrie rond de z- of t-waarde 0 is aangegeven met stippellijntjes.

U kunt nu naar hartenlust in cel C4 de lengte van de steekproef variëren, neem bijvoorbeeld een steekproef van vijftig waarnemingen en kijk wat er met de kansverdeling van de t-waarden gebeurt in de kolommen C, D en E.

4.6 Functies gebaseerd op de t-verdeling

In dit hoofdstuk hebben we op praktische wijze in Excel willen laten zien dat de link tussen populatiekengetallen en steekproefkengetallen in het geval van aselechte steekproeven verloopt via kansverdelingen. Daartoe werd een aselechte steekproef getrokken om een uitspraak over de populatie te doen en met de juiste kansverdeling in de hand kunnen we dan op basis van het steekproefgemiddelde iets zeggen over het populatiegemiddelde. U hoeft niet steeds de gehele verdeling erbij te betrekken, maar u dient wel te weten wat de kans is op zeer extreme waarden gegeven het gemiddelde en standaarddeviatie. Om dergelijke vragen te beantwoorden zijn allerlei

functies in Excel beschikbaar. Aangezien uw steekproef doorgaans zal bestaan uit een of twee schoolklassen – dus steekproeven met een relatief kleine omvang - bespreken we hieronder drie functies die betrekking hebben op de t-verdeling.

T.VERD() deze functie is hierboven uitgebreid aan de orde gekomen. Standaard geeft T.VERD(t;vrijheidsgraden;1) de eenzijdige overschrijdingskans van t (dus de kans dat t een bepaald getal is of nog extremer). Hieronder staan verschillende voorbeelden voor evalueren van verdeling bij $X=1,5$ of $-1,5$.

De tegenhanger van T.VERD() is de functie T.INV(). De T.INV(kans;vrijheidsgraden) functie geeft u de waarde van X terug bij een bepaalde overschrijdingskans. Deze wordt standaard tweezijdig uitgevoerd. Als u dus kans van 0,05 invoert dan geeft de functie u die waarde terug waarvan de linkeroverschrijdingskans én de rechteroverschrijdingskans allebei 0,025 zijn. Dus '=T.INV(0,05;19)' geeft de kritieke waarde in de t-verdeling met 19 vrijheidsgraden met een rechteroverschrijdingskans van $0,05 / 2$.

Bestudeer de voorbeelden in Tabel 4.16 goed. De lengte van onze steekproef is steeds 20 (dus het aantal vrijheidsgraden is 19). Daarnaast kunt u natuurlijk altijd nog in Excel zelf onder Help de nodige aanvullende informatie over deze en talloze andere functies opzoeken. Indien u een echt grote steekproef heeft getrokken kunt u over stappen op de z-verdeling.

	GEGEVEN	berekend		
	t	kans		
rechteroverschrijdingskans	1,50	0,0750	=T.VERD(1,5;19;1)	
linkeroverschrijdingskans	-1,50	0,0750	=T.VERD(ABS(-1,5);19;1)	=T.VERD(1,5;19;1)
linkeroverschrijdingskans	1,50	0,9250	=1-T.VERD(1,5;19;1)	
rechteroverschrijdingskans	-1,50	0,9250	=1-T.VERD(ABS(-1,5);19;1)	=1-T.VERD(1,5;19;1)
tweezijdige overschrijdingskans	1,50	0,1500	=T.VERD(1,5;19;2)	
	kans	t		
Kritieke waarde bij gegeven kans	0,0500	2,09	=T.INV(0,05;19)	
Kritieke waarde bij gegeven kans	0,1000	1,73	=T.INV(0,10;19)	
Kritieke waarde bij gegeven kans	0,1500	1,50	=T.INV(0,15;19)	

Tabel 4.16. Voorbeelden van de functies T.VERD() en T.INV().

In Tabel 4.16 veronderstellen we bij gebruik van de functie T.VERD() dat u al weet wat de t-waarde is (in ons voorbeeld is dat 1,5 of -1,5). Vaak zult u echter deze t-waarde zelf moeten berekenen op basis van uw steekproef. Uiteraard kunt u dit met een functie doen in Excel. Dit is de functie T.TOETS().

De functie T.TOETS() kent vier argumenten. De eerste twee argumenten hebben betrekking op de cellen waar de steekproefgegevens staan. De laatste twee argumenten beschrijven de verschillende veronderstellingen op basis waarvan u kunt gaan toetsen. Dit is samengevat in Figuur 4.17. Een leraar heeft twee proefwerken afgenomen. De resultaten van dit proefwerk staat in de kolommen B en

C. De kengetallen gemiddelde en standaarddeviatie per proefwerk heeft ze opgevraagd in de cellen E2:F4 van Tabel 4.17.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	cel A1	P1	P2		P1	P2									
2	II_1	7	7	gemiddelde	6,70	7,25									
3	II_2	7	7	stdev	1,72	1,48									
4	II_3	8	8												
5	II_4	9	9	0,0225	=T.TOETS(B2:B21;C2:C21;1;1)										
6	II_5	4	5	0,1427	=T.TOETS(B2:B21;C2:C21;1;2)										
7	II_6	8	9	0,1428	=T.TOETS(B2:B21;C2:C21;1;3)										
8	II_7	6	7												
9	II_8	7	5	0,0450	=T.TOETS(B2:B21;C2:C21;2;1)										
10	II_9	8	7	0,2855	=T.TOETS(B2:B21;C2:C21;2;2)										
11	II_10	7	6	0,2856	=T.TOETS(B2:B21;C2:C21;2;3)										
12	II_11	8	10												
13	II_12	3	5												
14	II_13	6	7												
15	II_14	7	8												
16	II_15	8	8												
17	II_16	3	6												
18	II_17	5	6												
19	II_18	8	9												
20	II_19	8	9												
21	II_20	7	7												
22															

Tabel 4.17. Zes mogelijke t-toetsen op dezelfde twee steekproeven.

Elke t-toets die de lerares in Tabel 4.17 heeft berekend, is gebaseerd op gedachten of veronderstellingen in het hoofd van de lerares. Haar theorie bepaalt deels welke variant van de t-toets ze kan gebruiken. Stel, de lerares heeft de kinderen leren rekenen op basis van het realistisch rekenen en dat afgerond met een (eerste) proefwerk. Inmiddels heeft ze gelezen dat oefening kunst baart in het rekenonderwijs en op grond daarvan heeft ze de methode realistisch rekenen met veeltalige uitleg en voorbeelden vervangen door meer abstracte sommen te laten maken. Ze rondt ook deze periode af met een (tweede) proefwerk. Ze meent nu inderdaad dat de kinderen beter kunnen rekenen nadat ze vele abstracte oefeningen gemaakt hebben en verwacht dan ook dat proefwerk 2 veel beter gemaakt wordt. Vervolgens berekent ze de kans dat de cijfers op proefwerk 2 hoger zijn dan proefwerk 1 onder de verwachting dat er geen verschil zou zijn tussen beide methoden. Dus ze gebruikt de eenzijdige t-toets: '= T.TOETS(...;...;1;...)'.

Het kan ook zijn dat ze weliswaar de discussie over het rekenonderwijs heeft gevolgd, maar niet zeker is van de argumenten van Van de Craats ([www. staff.science.uva.nl/~craats](http://www.staff.science.uva.nl/~craats)). Collega's overtuigen haar van het feit dat de benadering van Van de Craats helemaal niet beter of slechter is. De lerares gaat nu rekenen onder de verwachting dat er geen verschil is tussen scores op proefwerk 1 en proefwerk 2. Ze laat in het midden of proefwerk2 juist beter of slechter gemaakt kan worden dan proefwerk 1. Ze toetst dan tweezijdig '= T.TOETS(...;...;2;...)'.

Het laatste argument van de functie T.TOETS(...;...;...;1) heeft te maken met de vraag of ze de verschillen tussen proefwerk 1 en proefwerk 2 koppelt aan individuele leerlingen. Dus ze vergelijkt per leerling de score op proefwerk 1 en proefwerk 2 (t-toets voor gepaarde waarnemingen, ook wel de t-toets voor verschillscores genaamd). Of ze kijkt enkel naar de gemiddelden per proefwerk. Hierbij kan ze dan ook nog eens de schatter voor de standaarddeviatie per proefwerk apart meewegen '= T.TOETS(...;...;...;3)', of ze kan besluiten de varianties van beide steekproeven te poolen (dus gewoon bij elkaar voegen) '= T.TOETS(...;...;...;2)'.

5 Toetsen

Bij toetsen gaat het steeds om een steekproef en de vraag is wat deze steekproef ons kan zeggen over de populatie. Dit principe moet u goed in de gaten houden. Op basis van de getrokken steekproef worden kengetallen vastgesteld: vaak, maar lang niet altijd, zijn dat het gemiddelde en de standaarddeviatie. Bij het trekken van een steekproef speelt kans een grote rol. De effecten van een dergelijk kansmechanisme moeten we verdisconteren ten einde uitspraken te kunnen doen op grond van onze steekproef over de populatie. Immers, de ene keer zit een leerling wel in de steekproef, en de volgende keer misschien niet. Het verdisconteren van dit kansmechanisme gebeurt door te werken met zogenaamde theoretische kansverdelingen. Er zijn heel erg veel theoretische kansverdelingen. De belangrijkste theoretische waarmee we in het onderwijs te maken hebben zijn waarschijnlijk de standaard normale, de t- verdeling en chikwadraatverdeling.

In dit hoofdstuk worden vier toetssituaties uit ons boek 'Klassenonderzoek? Dat doe je zól' uitgewerkt. Dat zijn achtereenvolgens: (1) t-toets voor het gemiddelde, (2) t-toets voor het verschil tussen twee gemiddelden, (3) t-toets voor verschilscores en de (4) chikwadraattoets voor kruistabellen. Voor het beantwoorden van deze vragen volgen we de algemene aanpak zoals beschreven door Brinkman.¹³ Dit schema bestaat uit zes verschillende stappen.

- Stap 1 en 2: formuleer achtereenvolgens de nulhypothese en alternatieve hypothese;
- Stap 3: kies een onbetrouwbaarheidsniveau α . Dat is de kans op een fout van de eerste orde die u bereid bent om te accepteren. Een fout van de eerste orde wordt gemaakt als u de nulhypothese verworpt terwijl die toch waar is;
- Stap 4: kies de juiste toets en bereken de toetsingsgrootte. Vervolgens wordt de kans op deze toetsingsgrootte vastgesteld onder de veronderstelling dat de nulhypothese waar is. Deze kans heet p-waarde;
- Stap 5 en 6: vergelijk tot slot de gevonden p-waarde met de gekozen α .

Als de p-waarde $< \alpha$ dan wordt de nulhypothese verworpen en de alternatieve hypothese geaccepteerd. U bent dan van mening dat de kans op een extreme uitkomst in de steekproef niet meer afgedaan kan worden als 'steekproeftoeval'. Er is iets bijzonders aan de hand en in dit geval wordt gesproken van statistisch significant.

Als de p-waarde $> \alpha$ dan accepteert u de nulhypothese. De extreme uitkomst is kennelijk niet extreem groot of klein en is 'af te doen' als steekproeftoeval.

De vier hieronder besproken toetsingssituaties worden steeds volgens dit schema doorgenomen. De moeilijkste stap is in feite het formuleren van de nul- en alternatieve hypothese zijn. Ook het kiezen van de juiste toets is vaak lastig. We hebben dit laatste geprobeerd op te lossen door vier expliciet voorkomende toetsingssituaties in de onderwijspraktijk te benoemen.

¹³ Brinkman, J. (2001). *Cijfers spreken. Statistiek en methodologie voor het hoger onderwijs*. Groningen: Wolters Noordhoff.

5.1 Toetsingssituatie 1: t-toets voor het gemiddelde

Is een (steekproef-)gemiddelde groter dan (kleiner dan, verschillend van) een bepaalde (theoretische) waarde?

Stel, een docent leest in een testhandleiding dat het algemeen gemiddelde intelligentieniveau van VWO-leerlingen gelijk is aan 115. Hij vermoedt dat vanwege een selectief toelatingsbeleid het gemiddelde intelligentiepeil van de VWO-leerlingen aan zijn school, het grote Alblasserwaardcollege, hoger is dan 115. De docent wil dus aantonen dat zijn leerlingen hoger scoren dan het populatiegemiddelde.

- Stap 1 en 2:

De docent veronderstelt dat het populatiegemiddelde 115 bedraagt, dat staat immers in de testhandleiding. In symbolen: $H_0 : \mu \leq 115$. Alle andere uitkomsten die afwijken van wat onder de nulhypothese zijn gespecificeerd, vormen tezamen de alternatieve hypothese (H_a of H_1). In symbolen geschreven ziet dit er als volgt uit: $H_1 : \mu > 115$. Het volgende paar hypothesen wordt getoetst:

$$H_0 : \mu \leq 115 \text{ versus } H_1 : \mu > 115$$

- Stap 3:

Het onbetrouwbaarheidsniveau α stellen we vast op 0,05. Dat wil zeggen dat als we een kans vinden van 5% of hoger op de geëigende toetsingsgrootte, dan is er geen aanleiding om de nulhypothese te verwerpen.

- Stap 4:

De docent trekt een steekproef in zijn eigen klas en laat 36 leerlingen een IQ test afnemen. Het steekproefgemiddelde bedraagt 118 met een standaarddeviatie van 14. De vraag is nu wat de kans is dat hij een steekproefgemiddelde vindt van 118 (of nog groter) onder de aanname dat in de populatie het gemiddelde 115 (of nog lager is). Op basis van de steekproef wordt de toetsingsgrootte vastgesteld en vervolgens de kans op deze toetsingsgrootte opgezocht onder de aanname dat H_0 juist is.

$$\text{t-waarde: } t = \frac{\bar{x} - \mu}{s_x / \sqrt{n}} = \frac{\bar{x} - \mu}{s_x} * \sqrt{n} = \frac{118 - 115}{14} * \sqrt{36} = 1,29.$$

De kans dat $t \geq 1,29$ onder aanname dat H_0 juist is, is niets anders dan de rechteroverschrijdingskans. Dus '=T.VERD(1,29;35;1)' = 0,1028. De p-waarde is 10%.

- Stap 5 en 6:

Deze kans dat we een t-waarde vinden van 1,29 onder aanname dat populatiegemiddelde 115 is bedraagt 10%. En dat is best groot. Het is in ieder geval groter dan 5% (=onbetrouwbaarheidsniveau α) De redenering is nu als volgt: Het is dus heel wel mogelijk wanneer het gemiddelde van de populatie 115 is, een steekproef te trekken met een gemiddelde van 118. We handhaven de nulhypothese want we hebben niet aangetoond dat het gemiddelde van de vwo-leerlingen op het Alblasserwaardcollege echt hoger is dan 115.

5.2 Toetsingssituatie 2: t-toets voor het verschil tussen twee gemiddelden

Zijn twee gemiddelden verschillend (of: is het ene gemiddelde groter dan het andere gemiddelde)?

Een conrector heeft als theorie dat aan haar school, het Plasterkcollege, de atheneumleerlingen assertiever zijn dan de havo-leerlingen. Zij gaat dit onderzoeken door middel van een steekproef onder tien atheneum- en tien havo-leerlingen. Uiteraard formaliseert zij haar theoretische uitgangspunten.

- Stap 1 en 2:

Het volgende paar hypothesen wordt getoetst:

$$H_0 : \mu_{\text{atheneum}} \leq \mu_{\text{havo}} \text{ versus } H_1 : \mu_{\text{atheneum}} > \mu_{\text{havo}}$$

- Stap 3:

Het onbetrouwbaarheidsniveau α stellen we vast op 0,05. Dat wil zeggen dat als we een kans vinden van 5% of hoger op de geëigende toetsingsgrootte, dan is er geen aanleiding de nulhypothese te verwerpen.

- Stap 4:

Ze neemt aan een steekproef leerlingen (tien atheneumleerlingen en tien havo-leerlingen) een assertiviteitsvragenlijst af. De scores staan in Tabel 5.1.

	A	B	C	D	E
1	Scores assertiviteitstest (n=20).			Cel D1	
2		havo	atheneum		
3	leerling 1	81	88		0,042139291
4	leerling 2	92	96		
5	leerling 3	72	92		
6	leerling 4	86	80		
7	leerling 5	85	114		
8	leerling 6	108	100		
9	leerling 7	88	96		
10	leerling 8	80	107		
11	leerling 9	101	96		
12	leerling 10	96	103		
13					
14	Gemiddelde	88,90	97,20		
15	Standaarddeviatie	10,66	9,61		

Tabel 5.1. Scores op assertiviteitstest onder tien atheneum- en tien havo-leerlingen.

Voor het overzicht hebben we ook maar even de beide steekproefgemiddelden en standaarddeviaties opgevraagd met de functies 'GEMIDDELDE()' EN 'STDEV()'. Dat moet u altijd doen voor uzelf. U overziet dan veel beter wat de kern van de zaak is: de atheneumleerlingen scoren inderdaad assertiever op de gebruikte test. Als dat niet zo was, dus als de havo-leerlingen gemiddeld hoger hadden gescoord dan de jongens en meisjes op het atheneum, dan was er niet eens iets om te toetsen.

De kansverdeling van de toetsingsgrootte kunt u nu opvragen met behulp van de functie `=T.TOETS(B3:B12;C3:C12;1;3)` en de uitkomst is 0,0421 (zie cel E3). Eerst even de argumenten nalopen in deze functie. We tellen drie keer een ';' dus er zijn vier argumenten. Het eerste argument verwijst naar de steekproefresultaten van de havo-leerlingen (B3:B12). Het tweede argument geeft de scores op het atheneum weer (C3:C12). Het derde argument, het getal 1 zegt dat u de toets eenzijdig uitvoert, kijk dit even na onder Stap 1 en 2! En het vierde argument vraagt u om aan te geven wat voor soort steekproef het is. Het getal 3 geeft aan dat het twee steekproeven met ongelijke varianties betreft.

- Stap 5 en 6:

De kans op een verschil van 8.30 (=97,20-88,90) tussen beide steekproefgemiddelden onder de aanname dat het gemiddelde van beide groepen gelijk zou zijn, bedraagt 4%. Aangezien de p-waarde (4%) < α (5%) wordt H_0 verworpen ten gunste van H_a . Atheneumleerlingen zijn inderdaad assertiever en de visie van docent wordt bevestigd door haar eigen onderzoek.

5.3 Toetsingssituatie 3: t-toets voor verschillscores

Zijn de scores van vóór de treatment verschillend van de scores van ná de treatment (of: zijn de scores van vóór de treatment groter of lager)?

Een steekproef van 15 leerlingen van het Abe Lenstracollege in Herenveen maakt een rekentoets. Daarna krijgen zij extra rekenlessen. Na afloop van deze extra rekenlessen wordt dezelfde rekentoets weer afgenomen. De scores zijn de tweede keer wat hoger dan de eerste keer. In deze situatie toetsen we dezelfde groep leerlingen dus twee maal: vóór en ná de rekenlessen. We spreken dan van twee afhankelijke steekproeven: afhankelijk in de zin van als u een leerling de eerste keer een toets afneemt, moet u hem of haar ook de tweede keer een toets afnemen. Hoewel het dezelfde groep leerlingen betreft, spreken we in de statistiek toch van twee steekproeven: we verwachten dat de groep qua rekenprestatie de tweede keer anders is (hoger scoort), dus het is de tweede keer een 'andere' steekproef. Zo spreken we hier ook van twee populaties: de populatie leerlingen van vóór de rekenlessen en de populatie (dezelfde) leerlingen van ná de rekenlessen.

- Stap 1 en 2:

Het volgende paar hypothesen wordt getoetst:

$$H_0 : \mu_{\text{voor}} \geq \mu_{\text{na}} \text{ versus } H_1 : \mu_{\text{voor}} < \mu_{\text{na}}$$

Dit kan men ook geschreven als:

$$H_0 : \mu_{\text{voor}} - \mu_{\text{na}} \geq 0 \text{ versus } H_1 : \mu_{\text{voor}} - \mu_{\text{na}} < 0$$

- Stap 3:

Het onbetrouwbaarheidsniveau α stellen we vast op 0,05. Dat wil zeggen dat als we een kans vinden van 5% of hoger op de geëigende toetsingsgrootte, dan is er geen aanleiding de nulhypothese te verwerpen.

- Stap 4:

Hij neemt aan twee keer dezelfde groep van vijftien leerlingen een test af. Dat is de steekproef. De scores staan in Tabel 5.2 en het is nu zaak om de toetsingsgrootheid te bepalen.

	A	B	C	D
1		scores op eerste toets A	scores op tweede toets B	Cell D1
2	leerling 1	18	19	
3	leerling 2	14	15	
4	leerling 3	9	8	0,0113
5	leerling 4	12	13	
6	leerling 5	15	15	
7	leerling 6	8	12	
8	leerling 7	9	11	
9	leerling 8	10	9	
10	leerling 9	11	12	
11	leerling 10	9	10	
12	leerling 11	16	17	
13	leerling 12	12	13	
14	leerling 13	9	8	
15	leerling 14	7	11	
16	leerling 15	8	9	
17				
18	Gemiddelde	11,13	12,13	
19	Standaarddeviatie	3,29	3,27	

Tabel 5.2. Scores van vijftien leerlingen op twee toetsen voor (A) en na behandeling (B).

Ook nu begint u met gewoon het berekenen van het gemiddelde en de standaarddeviatie van de beide steekproeven. De toets voor verschillcores staat ook bekend onder de naam t-toets voor gepaarde waarnemingen. U vraagt deze op: '=T.TOETS(B2:B16;C2:C16;1;1)' = 0,0113. Het laatste argument in deze functie (de laatste '1' dus) geeft aan dat het om een gepaarde steekproeven gaat.

- Stap 5 en 6:

Onder de aanname dat er geen verschil zou zijn tussen vóór (A) en ná de behandeling (B) blijkt de kans op een dergelijke 'extreme' uitkomst erg klein te zijn: 0,0113. Een p-waarde van $0,0113 < 0,05$ en dus verwerpen we de nulhypothese. De extra rekenlessen sorteren wel degelijk effect.

U kunt overigens dit probleem ook op een geheel andere manier oplossen. Die manier wordt in ons boek "Klassenonderzoek? Dat doe je zól!" gevolgd. In het kort komt het neer op het volgende. U definieert het verschil tussen de individuele scores op A en B. Deze verschillscore zet u om in een t-

waarde met $15-1=14$ vrijheidsgraden. In feite heeft u dan toetsingssituatie 3 omgewerkt naar toetsingssituatie 1: de t-toets voor het gemiddelde.

5.4 Toetsingssituatie 4: chikwadraattoets voor kruistabellen

Eerder hebt u in dit KG kennis gemaakt met de Griekse letters μ en σ . We gaan voor de vierde toets een andere Griekse letter introduceren: χ (spreek uit: chi). U weet dat Griekse letters gebruikt worden om kengetallen op populatieniveau aan te duiden. Een extra bijkomstigheid is wel dat we werken met het kwadraat van dit kengetal. De populatieparameter is dus niet χ (chi), maar χ^2 (chikwadraat). Dit is minder vreemd dan u misschien geneigd bent te denken: de inmiddels bekende σ symboliseert de standaarddeviatie op populatieniveau, terwijl σ^2 (eveneens een kwadraat!) niets anders is dan de variantie op populatieniveau. Dat we spreken over chikwadraat en niet over chi heeft te maken met de formule die gebruikt wordt voor het berekenen voor deze populatieparameter.

Een studiecoördinator aan een grote universiteit in het midden van het land vermoedt dat het stemgedrag het van studenten samenhangt met een groot aantal zaken waaronder de gevolgde studierichting. Het gaat hier om twee variabelen waarvan de antwoordalternatieven een nominale schaal vormen. Hij wil geen uitspraak over een mogelijk causaal verband tussen beide variabelen. Het is niet van belang of studierichting leidt tot sympathie voor een bepaalde politieke partij of dat studiekeuze juist het gevolg is van politieke gezindte. Kortom, de te onderzoeken vraag luidt of er sprake is van samenhang tussen studierichting en politieke gezindte; meer niet.

- Stap 1 en 2:

De nul- en alternatieve hypothese worden eerst in woorden uitgeschreven. Dat kan geen kwaad: al is het alleen maar om uw gedachten te expliciteren.

H_0 : Onder universitaire studenten is er sprake van samenhang tussen studierichting en politieke voorkeur.

versus

H_1 : Onder universitaire studenten is er geen sprake van samenhang tussen studierichting en politiek voorkeur.

Dit kan formeel geschreven worden als:

$$H_0: \chi^2 = 0 \text{ versus } H_1: \chi^2 > 0 .$$

We toetsen dus tweezijdig.¹⁴ Er is wel verband tussen beide nominale variabelen, maar de richting ervan laten we in het midden. Dus of er juist heel veel pedagogen (zeg maar onevenredig veel) PvdA kiezen of juist heel weinig (dus onevenredig weinig), laten we in het midden.

- Stap 3:

Het onbetrouwbaarheidsniveau α stellen we vast op 0,05. Dat wil zeggen dat als we een kans vinden van 5% of hoger op de geëigende toetsingsgrootte, dan is er geen aanleiding de nulhypothese te verwerpen.

- Stap 4:

¹⁴ In Excel staat een storende tekstuele fout. Als u de functie CHI.KWADRAAT() via het scherm Functie invoegen opvraagt, wordt vermeld dat CHI.KWADRAAT() de eenzijdige kans van de chikwadraatverdeling uitrekent. Dat is niet correct: dat moet **tweezijdig** zijn. De functie als zodanig berekent ook daadwerkelijk de tweezijdige overschrijdingskans.

Studierichting en politieke voorkeur van 382 studenten in de steekproef is opgenomen in Tabel 5.3.

	A	B	C	D	E	F	G
1		PvdA	VVD	SP	CDA	Totaal	cel G1
2	Sociologen	35	16	19	21	91	
3	Pedagogen	14	41	20	30	105	
4	Psychologen	12	19	37	25	93	
5	Geografen	16	20	18	39	93	
6	Totaal	77	96	94	115	382	

Tabel 5.3. Kruistabel van studierichting en politieke partij (n=382).

Het wordt eentonig: ook in deze toetsingssituatie wordt de kans berekend op een bepaalde toetsingsgrootte, gegeven dat de nulhypothese waar is. De aantallen in de cellen van de kruistabel onder H_0 worden berekend als het product van de randtotalen. Zoiets is redelijk eenvoudig met behulp van Excel te berekenen. We gaan dus eerst de tabel maken onder de aanname dat de nulhypothese waar is.

- Typ in cel A8: "Verdeling onder aanname H_0 is waar."
- Kopieer de hele tabel naar de cellen A9 tot en met F14. Bij het overnemen van Tabel 5.3 heeft u er misschien voor gekozen om de randtotalen niet over te typen, maar op te vragen met behulp van een functie. U heeft bijvoorbeeld in cel G2: '=SOM(B2:E2)', in cel B6 '=SOM(B2:B5)', et cetera getypt. Als u nu deze tabel in zijn geheel wilt kopiëren, dan kunt u dit weliswaar doen met <Ctrl>+<C> gevolgd door <Ctrl>+<V>. Het nadeel van deze handeling is dat de randtotalen dan nog steeds met behulp een functie genereert worden. Het leegmaken van de cellen A9:F14 resulteert dan in randtotalen die allemaal gelijk aan 0 zijn. U kunt dit oplossen door niet met behulp van <Ctrl>+<V> te plakken, maar door in de menubalk onder Bewerken de optie Plakken speciaal te kiezen. In het dialoogscherm Plakken speciaal selecteert u dan de radiobutton Waarden gevolgd door OK. In cel G2 staat dan niet langer '=SOM(B2:E2)' = 91, maar direct het getal '91'.
- Verwijder de inhoud van de cellen B10 tot en met E13. De randtotalen worden dus niet verwijderd.
- Typ in cel B10 '=F\$10*B14/\$F\$14' = 18,34.
- Kopieer cel B10 en plak dit over de cellen C10, D10 en E10.
- Typ in cel B11 '=F\$11*B14/\$F\$14' = 21,16 en voer soortgelijke berekeningen uit voor de overige cellen.
- typ in cel D16 de tekst 'controle'.
- Als laatste stap verifieert u of de som in de door u berekende cellen gelijk is aan 382. Typ in cel F16: '=SOM(B10:E13)' = 382.

Vervolgens hoeven we enkel nog de juiste toetsfunctie erbij te zoeken:

- typ in cel D 17 'chikwadraat'.
- typ in cel F17 '=CHI.TOETS(B2:E5;B10:E13)' = 0,0000 (op vier decimalen nauwkeurig). Het geheel staat in Tabel 5.4.

	A	B	C	D	E	F	G
1		PvdA	VVD	SP	CDA	Totaal	cel G1
2	Sociologen	35	16	19	21	91	
3	Pedagogen	14	41	20	30	105	
4	Psychologen	12	19	37	25	93	
5	Geografen	16	20	18	39	93	
6	Totaal	77	96	94	115	382	
7							
8	Verdeling onder aanname H0 is waar.						
9		PvdA	VVD	SP	CDA	Totaal	
10	Sociologen	18,34	22,87	22,39	27,4	91	
11	Pedagogen	21,16	26,39	25,84	31,61	105	
12	Psychologen	18,75	23,37	22,88	28	93	
13	Geografen	18,75	23,37	22,88	28	93	
14	Totaal	77	96	94	115	382	
15							
16				controle		382	
17				chikwadraat		0,0000	

Tabel 5.4. Kruistabel van studierichting en politieke partij (n=382) en de verwachte kruistabel onder de aanname dat nulhypothese juist is.

- Stap 5 en 6:

Dus de kans dat u een verdeling krijgt zoals in de echte steekproef van 382 studenten is wel heel erg klein als er inderdaad geen verband zou zijn tussen studierichting en politieke partijkeuze. Eigenlijk denken we dan dat de nulhypothese niet waar is. Het kan wel, maar erg waarschijnlijk is het niet! Formeel besluiten we dat $0,0000$ (p-waarde) $< 0,0500$ (alfa) en we verwerpen de nulhypothese en accepteren de alternatieve hypothese. Er is dus wel een verband tussen studierichting en politieke partijkeuze onder universitaire studenten.

Het is lastig dat we op het eind van de toets niet weten hoe groot de toetsingsgrootheid (χ^2) nu eigenlijk is. In een rapport wordt de waarde van de toetsingsgrootheid doorgaans wel netjes gerapporteerd. U kunt deze wel opvragen met hulp van de functie =CHI.KWADRAAT.INV(kans;vrijheidsgraden). Helaas is in ons voorbeeld de kans zo ontzettend klein dat Excel niet de waarde van χ^2 teruggeeft, maar de tekst '#GETAL'.